UNIVERSITY OF CALIFORNIA,
IRVINE

Leveraging Diversity to Improve the Wisdom of the Crowd

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Sciences

by

Lauren Elizabeth Montgomery

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

2024

DEDICATION

To my parents – *Alan and Jennifer* – who will always be home,
and to my siblings – *Adam, Alyssa, and Sydney* –
who are listed alphabetically and by birth order
but not, I might add, by who I love most although love them I do,
and to my best friend – *Kamya* – who has been family since the fourth grade.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank the US Air Force Research Laboratory's Continuous Learning Branch. Part of the work in this dissertation was enabled through my appointment to the Oak Ridge Institute for Science and Education (ORISE) Summer Research Internship Program. The views expressed in this dissertation are my own and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

I would like to thank my advancement committee – Michael D. Lee, Şule Güney, Megan Peters, Mark Steyvers, and Joachim Vandekerckhove – for their guidance, giving particular thanks to my dissertation committee – Michael D. Lee, Mark Steyvers, and Joachim Vandekerckhove – for taking the time to meet with me, answer my questions, and generally just nudge me along.

I want to extend a huge thank you to Clara Schultheiss and MyLee Ryan-Glass for answering my questions over the last five years and being invaluable sources of knowledge for how best to navigate through the administrative quagmire. I would like to thank all the members of the Bayesian Cognitive Modeling, aka the Lee Lab, for being willing to listen to my ideas and presentations over the years. The questions and camaraderie were both indispensable. A special thanks to Charles M. Baldini for all his help with the *Where's Waldo, Ohio? Using cognitive models to improve the aggregation of spatial knowledge* project.

I would also like to thank all of the friends who have helped and supported me along the way – you have made this program all the more enjoyable. Thank you Helen Braithwaite for opening up your house and sharing your wonderful baking. Thank you Adriana Chávez de la Peña and Jaime Islas Farias for joining me for numerous nights of tv watching. Thank you Angela Shen and Prawit Siriwat for always being down for tea and noodles no matter the when. Thank you Paulina Silva, Jeff Coon, and Sharanya Bashyam for writing with me even when no one else showed up. Thank you Nidhi Banavar, Nora Harhen, and Priyam Das for your delightful company during work and teatime. Thank you Mahbod Mehrvarz for your constant supply of organized nonsense, and thank you Talía Roman for our fun food adventures and drives to and from LA. Thank you Alexander Etz for your insightful questions and general statistics knowhow. Thank you Robyn Lin, Heidi Waite, and Sarah Schlund for all the ways you each made our apartment feel like home. Thank you Constantijn van der Poel, my parents, my siblings, and Kamya Gopal for always just being there for me, lending your continual support, and cheering me on at every step.

Thank you Holly Westfall for leading our lab adventures, sharing stickers, and jokes all while answering questions and knowing exactly what needed to be done next. Thank you Jesus Manuel Villarreal Ulloa for your continual friendship and help while we worked through the program. I have not forgotten that you tattled on me, but I suppose that is balanced out by all the quali-tea time we have spent together. Thank you Michael Lee for being such a wonderful advisor, teaching me how not to utilize some words when others work better, and even attempting to indoctrinate me in the ways of cricket. I have not and

cannot properly convey the magnitude of my thanks to everyone, but thank you for all your guidance, patience, friendship, and everything really. I cannot imagine my time at UC Irvine without you all, nor would I want to.

VITA

Lauren Elizabeth Montgomery

EDUCATION

Doctor of Philosophy in Cognitive Sciences **2024**
University of California, Irvine *Irvine, CA*

Master of Science in Statistics **2022**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Decision Science, with Add. Majors **2018**
Carnegie Mellon University *Pittsburgh, PA*

RESEARCH EXPERIENCE

Graduate Researcher **2019–2024**
University of California, Irvine *Irvine, California*

AFRL Math Psychology Internship **Summer 2022**
Oak Ridge Institute for Science and Education (ORISE)

AFRL Human Machine CO-Learning Psychology **Summer 2021**
Oak Ridge Institute for Science and Education (ORISE)

Undergraduate Research Assistant **2017–2018**
Carnegie Mellon University *Pittsburgh, PA*

TEACHING EXPERIENCE

Teaching Assistant **2019–2024**
University of California, Irvine *Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Where's Waldo, Ohio? Using cognitive models to improve the aggregation of spatial knowledge **2024**
Computational Brain & Behavior

Expert and novice sensitivity to environmental regularities in predicting NFL games **2023**
Judgment and Decision Making

REFEREED CONFERENCE PUBLICATIONS

The wisdom of the crowd and framing effects in spatial knowledge **July 2022**
Proceedings of the 44th Annual Conference of the Cognitive Science Society

AWARDS & HONORS

WoMP Travel & Networking Award **2023**
Women of MathPsych, €500 of travel funding

Social Sciences Instructional Fellowship **2020**
Social Sciences School at the University of California, Irvine

ABSTRACT OF THE DISSERTATION

Leveraging Diversity to Improve the Wisdom of the Crowd

By

Lauren Elizabeth Montgomery

Doctor of Philosophy in Cognitive Sciences

University of California, Irvine, 2024

Professor Michael D. Lee, Chair

This dissertation addresses how contextualized expertise and task design can improve wisdom of the crowd estimates. The first two chapters apply the wisdom of the crowd to two related tasks that require spatial knowledge. The third chapter applies the wisdom of the crowd to a subset ranking task.

In Chapter 1, I investigate how framing effects impact the wisdom of the crowd. Participants selected tiles that either represented US states or African countries in two frames, present and absent. I constructed three wisdom of the crowd estimates: an unweighted average, a confidence-weighted average, and a wisdom of the crowd within estimate that combines an individual's responses across frames. I found that combining the estimates from the two frames resulted in an improved wisdom of the crowd estimate.

In Chapter 2, I build on the wisdom of the crowd application for a task that again requires spatial knowledge. Participants supplied a point estimate and a radius centered at that point estimate for where various US cities were located. Unweighted and radius-weighted wisdom of the crowd estimates were more accurate than most individuals, but the cognitive model-based wisdom of the crowd estimates tended to be even more accurate. I describe how using cognitive modeling that contextualizes expertise led to improved wisdom of the crowd estimates.

In Chapter 3, I present a new extension for the Thurstone model to partial ranking data. Ranking tasks have usually had participants rank all items, but I present two different types of partial ranking tasks where either an experimenter or a participant selects the items to be ranked. I demonstrate how the Thurstone model can be used to generate wisdom of the crowd estimates, and speculate how other partial ranking tasks can be developed to better elicit diverse estimates from the crowd.

In all, these chapters detail specific applications of the wisdom of the crowd effect that better contextualize expertise, elicit multiple meaningful estimates from the same individual, and improve diversity. These methods are used in conjunction with cognitive modeling to produce improved wisdom of the crowd estimates.

INTRODUCTION

The wisdom of the crowd effect describes when group aggregation results in a more accurate estimate than that of a random individual group member (Galton, 1907; Surowiecki, 2004). Aggregating a group's estimates or judgments amplifies the shared signal in their responses while canceling out the idiosyncratic noise. These group aggregations can take on a variety of forms, but are most typically measures of central tendency, e.g., an arithmetic mean. While there are a vast number of applications for the wisdom of the crowd, there are several prerequisites for the wisdom of the crowd finding to hold: the crowd needs to be diverse, estimates or judgments need to be produced independently, and estimates or judgments should be elicited in a decentralized way (Surowiecki, 2004, p. XVIII). Furthermore, the wisdom of the crowd should be limited to situations where performance can be evaluated, either by knowing the ground truth or being able to determine it at a later time. There are other topics in the literature, like that of cultural consensus theory (Anders & Batchelder, 2012), that would be better references when evaluation is not possible or not an appropriate way to judge the accuracy of the aggregated estimate. Even with all these caveats, the wisdom of the crowd effect is versatile and particularly useful when the correct answer is difficult or costly to acquire, i.e., prediction or forecasting (Atanasov et al., 2017; Budescu & Chen, 2014; Da & Huang, 2019; Davis-Stober et al., 2015).

Considering how useful the wisdom of the crowd effect is, one general goal in the literature is to find how to aggregate the available information to get the *best* wisdom of the crowd estimate. Achieving this goal requires knowing both how to best elicit the existing wisdom from the crowd and how to improve the wisdom of the crowd estimates. With respect to the first question of how to elicit the crowd's knowledge most effectively, there is a strong emphasis on crowd construction and using multiple estimates from the same individual. Crowd construction, which is the focus of the wisdom of select crowds, tends to focus on winnowing the crowd down to just those who are "expert" enough. The wisdom of select crowds is specifically focused on finding smaller groups that concentrate particular measures of expertise while still demonstrating a wisdom of the crowd effect, specifically by quantifying the relative expertise of individuals in the crowd in some way (Mannes et al., 2014). Though these selected crowds can outperform other simple group estimates, one issue with focusing on expertise is that it is not necessarily transferable. This means that groups chosen on the basis of expertise alone may not perform as well on another related task or question.

Knowing how to select a better crowd is only part of the answer though, as another consideration for extracting more wisdom from the crowd is whether or not the single estimate or judgment that was collected is all that an individual knows. It might be the case that individuals want to convey their uncertainty about how long the Nile is, or that they would be equally willing to say that North Dakota *or* Maine is further north than Texas. Thus, it may also be desirable to gather multiple estimates or different types of estimates from the same individual. The wisdom of the crowd within refers to the approach of using multiple estimates from the same individual for inner (individual) and outer (group) wisdom of the crowd estimates. This approach inherently relies on obtaining independent estimates from the same individual, otherwise aggregating these estimates would increase, instead of decrease, the systematic error (Herzog & Hertwig, 2009). A simple way to do this is to allow time to pass between eliciting the estimates, potentially on the scale of weeks. Another approach is to have participants "consider-the-opposite" response to encourage individuals to seek out or consider previously overlooked information or information that is inconsistent with their current beliefs (Herzog & Hertwig, 2009; Lord et al., 1984). The wisdom of select crowds and the wisdom of the crowd within both address the question of how the existing wisdom can be extracted from the crowd, although it still leaves the question of how to further improve the wisdom of the crowd aggregations.

Usually, these wisdom of the crowd aggregations are simple arithmetic means, medians, or modes. There are endless variations on these methods, such as, geometric means (Lorenz et al., 2011), corrected means and medians (Kao et al., 2018), trimmed means (Jose & Winkler, 2008; Kao et al., 2018; Stock & Watson, 2004), Borda counts (Steyvers et al., 2009), weighted averages (Lyon & Pacuit, 2013), or the suprisingly popular method (Prelec et al., 2017; Lee et al., 2018). Generally speaking, simpler wisdom of the crowd aggregations like arithmetic means do not account for confidence or expertise, but can be extended to weighted arithmetic means that do. These methods can be more accurate, but are still limited to group aggregations and cannot say anything about the individual or parameters of interest like expertise. In contrast, cognitive modeling-based wisdom of the crowd estimates can be more accurate, provide both individual- and group-level inferences, and simultaneously provide information about latent parameters like expertise. Cognitive modeling can also help with unpacking and interpreting the diversity within the crowd. Diversity can be though of as the information that is unique to an individual, and this can easily be lost when aggregating across individuals to get the collective group estimate. As mentioned though, cognitive modeling captures information about the individuals and can consider more carefully how individual expertise should be distilled in the model's estimate. Thus, cognitive modeling is an especially useful method in improving the wisdom of the crowd estimates.

This dissertation is focused on using cognitive modeling to improve the wisdom of the crowd estimates. The three chapters of this dissertation are particular applications of the wisdom of the crowd to tasks that require spatial knowledge (Chapters 1 and 2) and subset ranking (Chapter 3); however, all of these applications are concerned with optimizing diversity. Throughout this dissertation, I consider alternative ways to optimize diversity by considering expertise in a more contextualized manner to further improve the wisdom of the crowd estimates. I focus on familiarity and confidence by collecting additional demographic information and by introducing it into the task structure for Chapters 1 and 2. This body of work suggests that expertise is not a constant individual feature across items and is instead more nuanced. I also focus on manipulations to the experimental structure that can introduce or preserve diversity in individuals' responses in Chapters 1 and 3.

The first chapter of the dissertation focuses on applying the wisdom of the crowd to a task that requires spatial knowledge. This application of a task with spatial knowledge is different from other existing work as it directly asks participants to provide spatial estimates instead of scalar estimates, discrete choices, or rankings. In this task, participants were asked to identify the location of various US states or African countries. Participants selected tiles in a present framing (e.g., "Where is Pennsylvania located? Select as few states as possible, but be sure Pennsylvania IS in the states you select") and absent framing (e.g., "Where is Pennsylvania NOT located? Select as many states as possible, but be sure Pennsylvania IS NOT in the states you select"). I found evidence of a framing effect where participants across both US states and African countries were more confident in terms of how many states or countries they selected in the present than absent frame. I then constructed three wisdom of the crowd estimates: an unweighted wisdom of the crowd estimate, a confidence-weighted wisdom of the estimate, and a wisdom of the crowd within estimate that was justified by there being a framing effect. Overall, both the wisdom of the crowd and the wisdom of the crowd within estimates outperformed most of the individuals. These findings suggest that individuals can produce diverse estimates with alternate question frames, and that the wisdom of the crowd estimates outperform individuals.

The second chapter of the dissertation focuses again on applying the wisdom of the crowd to a task that requires spatial knowledge, but goes further than Chapter 1 by using cognitive modeling to generate model-based wisdom of the crowd estimates. The experimental task had participants provide a point estimate of where a particular US city was located, and then using this point estimate as the center point produce a circle whose radius was certain to contain the target city. In their demographic information, participants were asked what US states they were familiar with which was operationalized as the states they had lived in or driven through frequently. I constructed an unweighted wisdom of the crowd estimate and a radius-weighted wisdom of the crowd estimate. These two statistical wisdom of the crowd estimates demonstrated the wisdom of the crowd effect. I then developed a series of cognitive models that included or excluded the radius judgments and assumed differences in individual expertise and city difficulty. These model-based estimates outperformed the statistical wisdom of the crowd estimates, as long as they assumed there was individual expertise. Model-based estimates were most accurate when they allowed for individual-by-city expertise. I replicated these findings using a dataset collected by Mayer & Heck (2023). In summary, the model-based estimates outperformed the statistical wisdom of the crowd estimates, and there is something about the individual-by-city expertise that should be further explored as it was not explained by familiarity with particular states as thought.

The third chapter of the dissertation deals with a ranking task rather than a task requiring spatial knowledge. Participants completed one of three task versions: complete ranking where they ranked the full set of items, experimenter-selected partial ranking where they ranked experimenter constructed subsets made at random or with respect to existing structure preexisting within the data, and individual-selected partial ranking where they selected with items they wanted to rank and only ranked their selected subset of items. The focus of these analyses was on the usage of a Thurstone model to produce a group aggregated ranking of the items. The inconsistencies between the different datasets limit the interpretation of results as they cannot be directly compared without caveat to each other. However, the cognitive model implementation of the Thurstone model to deal with partial ranking data is new and useful. Different experimental tasks can be developed to elicit diverse responses from the same individual, and these data can then be used to get at the underlying latent ranking.

I found that the Thurstone model produced reasonable, and in some cases very accurate, wisdom of the crowd estimates in addition to helpful information about latent parameters like an individual's inferred expertise. These findings demonstrate how the wisdom of the crowd can be applied to complete or partial ranking tasks aimed at preserving diversity in individuals' responses.

These chapters altogether extend the applications of the wisdom of the crowd and provide cognitive model-based approaches to further improve the wisdom of the crowd estimates. These results and methods consider alternative ways to optimize and preserve diversity in the crowd by considering expertise in a more contextualized manner and through different task designs. From this body of work, it cannot be assumed that expertise is a constant individual feature across items. Instead, we see evidence that expertise can be item specific, and so propose that cognitive models should consider breaking down expertise into component parts that are easier to identify that may help create diverse crowds with better contextualized expertise. This dissertation also explores several specific applications of the wisdom of the inner crowd where individuals are asked to generate multiple estimates, and by using the different question framings or task designs elicit multiple estimates from the same individual.

Chapter 1

The Wisdom of the Crowd and Framing Effects in Spatial Knowledge

Abstract

We study the wisdom of the crowd in the context of spatial knowledge, asking participants to identify US states and African countries on unlabeled tile maps. We use two question frames, asking participants to select where the target is present or eliminate where it is absent. Participants generally display overconfidence, often selecting small regions that do not include the target. We find strong wisdom of the crowd effects by aggregating participants' responses, especially by weighting the individual responses according to the size of their selection. The weighted crowd outperforms all but a few participants for the US states and all participants for the African countries. We also find wisdom of the crowd within effects, by aggregating the present and absent frames for the same participant. We discuss the implications of our findings for understanding how people express uncertain spatial knowledge and the potential use of crowd aggregation in real-world applications.

1.1 Introduction

The wisdom of the crowd is the finding that a crowd's aggregate judgment is more accurate than the judgment of a randomly sampled individual in the crowd (Galton, 1907; Davis-Stober et al., 2014; Surowiecki, 2004). Crowd superiority has been demonstrated in a range of contexts. The most common context is general knowledge, which examines the accuracy of answers to factual questions about geography, society, culture, entertainment, and other topics (Bennett et al., 2018; Lee et al., 2014; Prelec et al., 2017). Another context involves forecasting and predictions about political, social, sporting, and other events (Armstrong, 2001; Boon, 2012; Da & Huang, 2019; Klugman, 1947; Lee et al., 2018; Miller et al., 2012; Page & Clemen, 2013). A third context involves group settings in which individuals interact or compete with each other to generate judgments or estimates about stimuli (Atanasov et al., 2017; Christiansen, 2007; Lee et al., 2011b; Lyon & Pacuit, 2013; Ray, 2006). In all of these contexts, the required judgments can take different forms, including scalar estimates (Jenness, 1932; Farnsworth & Williams, 1936), discrete choice (Lee et al., 2018; Prelec et al., 2017), rank orderings (Bruce, 1935; Gordon, 1924; Knight, 1921; Lee et al., 2014; Miller et al., 2012), or sequential decisions (Thomas et al., 2021; Zhang & Lee, 2010).

In this study, we explore the wisdom of the crowd in the context of spatial knowledge by asking people to identify US states or African countries on unlabeled tile maps. Some previous research on spatial or geographical knowledge has focused on scalar estimates ("what is the height of Mount Everest?"), discrete choices ("is Reno east or west of San Diego?"), or rankings ("order the following US states from west to east") rather than direct spatial judgments. Other previous research has presented spatial targets and then required direct spatial judgments (Juni & Eckstein, 2017), although this type of task involves immediate perceptual rather than longer-term memory-based knowledge. The most relevant previous work studies how accurately people can identify locations on a map (Fu et al., 2017, 2020; Mayer & Heck, 2023). Our task involves people's memory for spatial knowledge and requires them to express that knowledge in a direct and detailed way by selecting a spatial region.

An interesting feature of our task is that it allows the same question to be framed in different ways. People are asked to identify a target US state or African country by selecting as many states or countries they need to be confident that the target is included in their set. We call this the present framing. They are also asked to identify a target state or country by indicating a set of states or countries that are *not* the target. We call this the absent framing. Being able to collect both of these judgments raises the issue of framing effects (Levin et al., 1998; Tversky & Kahneman, 1981) and, in particular, whether the inherent uncertainty in forming regions is managed differently between the frames. Previous research on elimination and inclusion, the same dichotomy that we use, suggests that using these frames will produce some non-complementarity in the generated choice sets (Shafir, 1993; Yaniv & Schul, 1997).

Asking multiple questions also allows us to consider the phenomenon known as the wisdom of the crowd within, in which multiple judgments from the same individual are aggregated. A basic challenge for the wisdom of the crowd within is that using only judgments from one individual results in correlated judgments, which limits the improvement in the aggregate. Accordingly, an effort is made to make the judgments as independent as possible. This has been achieved by increasing the time interval between estimates (Vul & Pashler, 2008) or having participants use various question framing strategies, such as consider-the-opposite (Herzog & Hertwig, 2009; Lord et al., 1984), starting from scratch (Herzog & Hertwig, 2014), or having the individual combine their previous estimates in some way (Herzog & Hertwig, 2009; Larrick & Soll, 2006). These question framing strategies work because the participant has to consider additional information or approach the question differently. In our spatial knowledge context, being able to ask about the location of targets in terms of presence and absence provides two natural contexts for asking the same individual about the same information.

The remainder of this paper is organized as follows. We first describe the experimental design and the framing effects on participants' judgments and how participants manage the uncertainty inherent in the task. To test for the wisdom of the crowd, we develop two approaches for aggregating crowd judgments and compare their performance to individual judgments. To test for the wisdom of the crowd within effects, we examine improvements in individual judgments resulting from aggregating their two judgments. Finally, we examine how the crowd aggregate improves as a function of the number of individuals in the crowd. We close with a discussion of our main results and directions for future research.

1.2 Experiment

1.2.1 Participants

50 participants were recruited using Prolific (Prolific, 2022) for each of the US states and African countries conditions in a between-participants design. All participants were current US residents and provided basic demographic information including their age, whether they attended high school in the US, and their familiarity with each of the US states or African countries.

1.2.2 Stimuli

Figure 1.1 shows the tile maps presented to participants in each trial. These are standard configurations used in data journalism.[1] The US states map was restricted to the 48 continental US states, and the African countries map was restricted to 51 of the 54 countries by excluding Comoros, Mauritius, and the Seychelles.

[1] See, for example, `https://blog.apps.npr.org/2015/05/11/hex-tile-maps.html` and `https://public.tableau.com/app/profile/neil.richards/viz/Malaria_14/Dashboard1`.

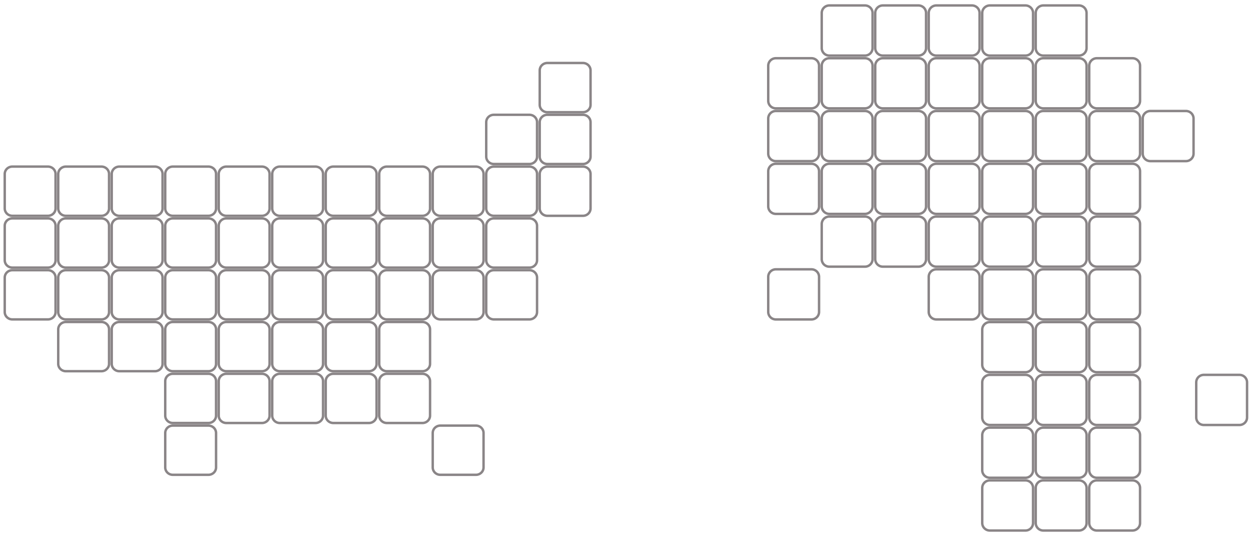

Figure 1.1: Tile map stimuli for the US states (left) and African countries (right) conditions.

The tile maps make responding to the task simple and responses easy to visualize. They also introduce some irreducible uncertainty because even participants with perfect geographical knowledge will still be uncertain about the exact translation between the true geography and the tile layout. For example, South Africa could reasonably be any of the three tiles at the bottom of the African map. Thus, when responding to the questions, participants need to consider both the uncertainty in their spatial knowledge and the uncertainty that the tile layout introduces.

1.2.3 Method

Every participant was given every state or country as a target on a trial in both the present and absent framings. The two framings were blocked so that all of the targets were presented in one frame before changing to the other. The order of the framings was randomized, as was the order of the targets.

In the present framing, participants were asked "Where is X located? Select as few states/countries as possible, but be sure X IS in the states/countries you select." In the absent

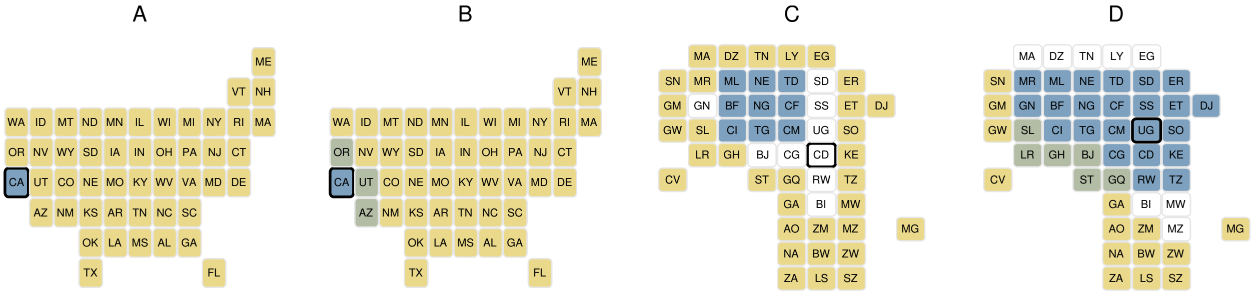

Figure 1.2: Four illustrative individual participant responses to particular target states and countries in both frames. States and countries selected in only the present frame are colored yellow, selected in only the absent frame are colored blue, selected in both frames are colored blue-yellow, and selected in neither frame are colored white.

framing, participants were asked “Where is X NOT located? Select as many states/countries as possible, but be sure X IS NOT in the states/countries you select.” Each question was answered sequentially with participants being asked not to look up any information but rely instead on their general knowledge and memory. Participants were not allowed to return to or view previous responses, and they did not receive any feedback. At the completion of all of the target questions in both frames, participants were asked for their demographic information.

1.3 Framing Effects and Managing Uncertainty

To analyze framing effects, we looked at how complementary the participants’ responses were. Complementary means that a participant’s response contained the same information in both frames. Figure 1.2 shows participant-level responses for both the present and absent frames for four illustrative cases. In each panel, the tile for the target state or country is outlined in black. The participant’s selections made only in the present frame are in blue, and their selections made only in the absent frame are in yellow. Tiles for states or countries selected in both frames are a blended blue-yellow color, and tiles selected in neither frame are white. This means that the extent of blue versus yellow regions indicates the confidence

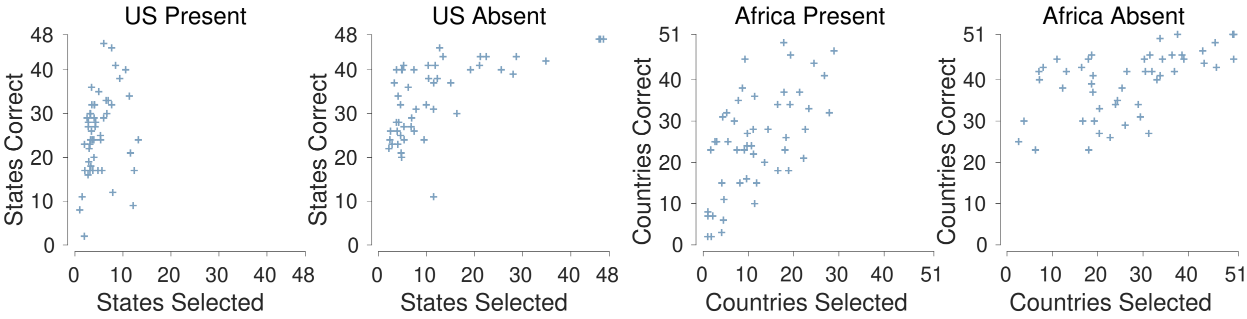

Figure 1.3: Individual participant performance in both conditions and both frames. Each blue cross corresponds to a participant, showing the average number of selections they made and the numbers of states or countries correctly included in their selections.

of the knowledge expressed by the participant. For example, the participant in panel A is very confident in locating California, the participant in panel C is less confident (and wrong) in locating the Democratic Republic of the Congo, and the participant in panel D has low confidence in locating Uganda.

The presence of blue-yellow and white tiles indicates that the participant's responses across the two frames are not perfectly complementary. In panel A, the participant made logically complementary selections for California, while in panel B the participant selected some states neighboring California in both the present and absent frames. This suggests the participant in panel B was less confident in the present than the absent frame. In contrast, the participant in panel C is more confident in the present frame and less confident in the absent frame. The participant in panel D is hard to characterize, since their present and absent frame responses are quite inconsistent, with some countries selected in both framings and others selected in neither. Participants rarely provided strictly complementary responses. On average, participants provided 5.7 complementary responses in the US states condition and 2.0 per country in the African countries condition. They had some overlap in 19.3 and 24.3 states and countries, respectively. They selected some tiles in neither frame in 38.4 and 45.9 states and countries, respectively.

Consistent with the task instructions, we measure a response as accurate if the target is

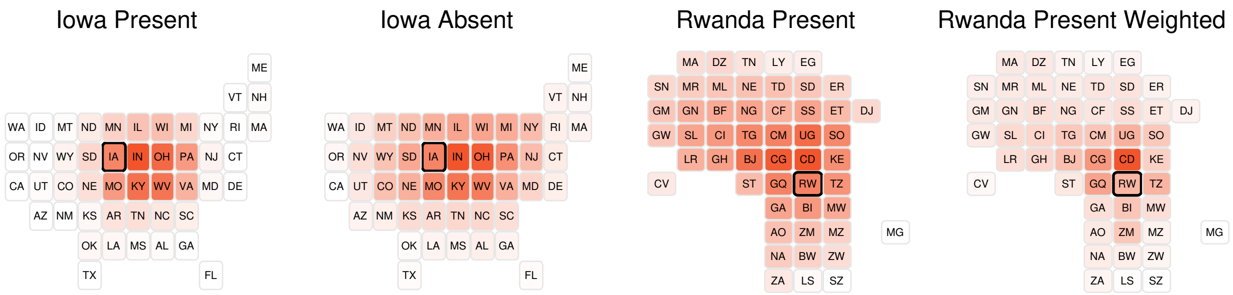

Figure 1.4: Examples of aggregate crowd responses. The two left-most panels show the unweighted proportion of participants who selected each state while targeting Iowa in the present and absent frames, with darker red colors indicating greater proportions. The two right-most panels show the unweighted and confidence-weighted proportion of participants who selected each country while targeting Rwanda in the present frame.

included in the participant's selections in the present frame or not included in their selection in the absent frame, regardless of the size of the regions they selected. Figure 1.3 shows the relationship between the number of states or countries selected and this measure of participant accuracy. The four panels correspond to the US states and African countries conditions and the present and absent frames. To allow direct comparisons between the two frames, participant responses in the absent framing have been inverted so that they indicate the states or countries the participant selected as including the target. This means that less confident behavior now consistently corresponds to higher numbers of selections and more confident behavior corresponds to lower numbers of selections.

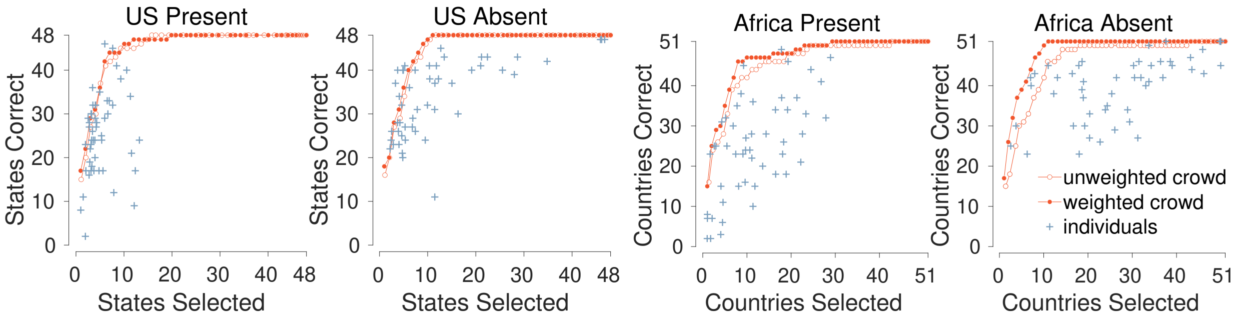

Figure 1.5: Individual participant and crowd performance in both conditions and both frames. Blue crosses correspond to participant performance. Red curves correspond to unweighted and weighted crowd performance, showing the average number of selections made and the numbers of states or countries correctly included.

The striking feature of Figure 1.3 is that very few participants achieve high levels of accuracy. This likely reflects both a lack of perfect knowledge and a failure to compensate by selecting enough states or countries. In the present frame, participants selected an average of 5.2 states and 11.8 countries, correctly including an average of 25.5 states and 25.0 countries. More selections are made in the absent frame, especially for US states. The average numbers selected are 12.4 states and 26.8 countries. These expanded selections lead to greater average accuracies of 33.8 states and 39.6 countries.

There is no reason, however, participants cannot achieve perfect accuracy in both frames. In fact, this is what the task instructions require. A participant who has little relevant geographical knowledge should select many of the states or countries in the present frame and few in the absent frame. No participants were completely accurate in the US states condition. The four participants who achieved complete accuracy in the African countries condition did so in the absent frame by eliminating very few countries. The fact that most participants achieve modest accuracy suggests that they are overconfident in their selections. The explanation cannot be as simple as wanting to avoid effort, since the way to achieve high accuracy in the absent frame is the least effortful. Most participants provide effortful responses in the absent frame that still exhibit overconfidence.

1.4 The Wisdom of the Crowd

The simplest way to form an aggregate crowd judgment is to count the proportion of times each state or country is selected by a participant. A more complicated method weighs the individual selections according to their confidence. A natural measure of confidence is the number of states or countries selected: that is, the number selected in the present frame and the number not selected in the absent frame. For example, if a participant selects 10 states, each of their selections will have 1/10th the value of a participant who just selected one state. Weighting individual judgments in this way implements the idea that more confident participants should have more influence on the crowd judgment (Lyon & Pacuit, 2013).

Figure 1.4 demonstrates these two approaches to crowd aggregation using heat map visualizations. The states and countries are shaded according to the aggregated group proportions. The left-most panels show the present and absent frames for the target state Iowa. It is clear that the crowd selection is more concentrated (less disperse) in the present frame, consistent with individuals making relatively fewer selections. The right-most panels show the unweighted and confidence-weighted crowd judgments for the target country Rwanda. The confidence-weighted aggregate is much more concentrated than the unweighted aggregate. This is a natural consequence of giving less weight to each selection made by participants who made many selections overall.

Crowd judgments are inherently graded and give a probability that each state or country is the target, unlike individual judgments in which every state or country is either selected or not selected. Accordingly, there is no natural single measure of crowd accuracy. Instead, there is a set of measures, depending on where the graded responses are thresholded. A simple way to set these thresholds is by ranking the probabilities and setting a threshold k so that only states or countries in the top-k are considered to be selected by the crowd. For example, if $k = 1$, the crowd response is the modal (most likely) state or country. In all four of the illustrative examples in Figure 1.4, this response would be incorrect. As the threshold is increased, to allow the top-two or top-three or more possibilities, the crowd will become more accurate at the expense of making more selections.

Figure 1.5 superimposes crowd performance on the individual performance shown in Figure 1.3. The red curves correspond to crowd performance, starting with the modal response and ranging to increased numbers of selections and accuracy (the non-integer values for selections are the result of ties in probabilities). These curves are shown for both the unweighted and confidence-weighted crowds. Better performance corresponds to small numbers of selec-

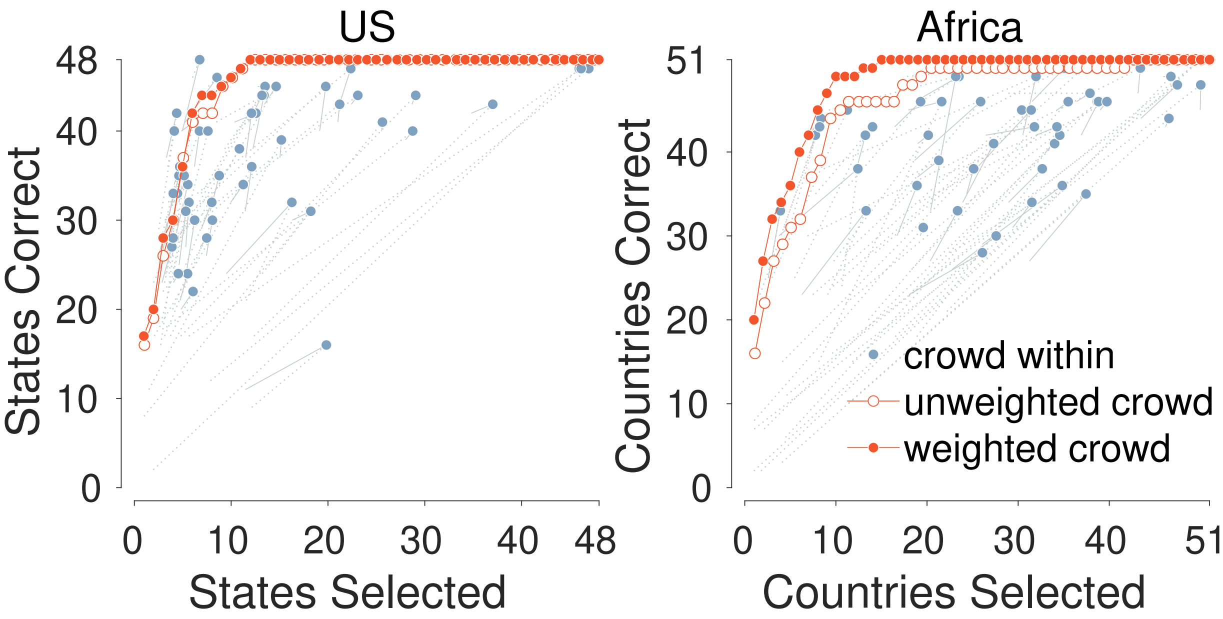

Figure 1.6: Wisdom of the crowd within performance for both conditions. Blue markers show aggregate performance across both frames for individual participants with lines connecting to their performance in each frame. The red curves show the the unweighted and confidence-weighted crowd performance.

tions with high accuracy. The unweighted and confidence-weighted curves are very similar in the US states condition but the weighted crowd clearly performs better in the African countries condition, especially for the absent frame.

Comparing crowd and individual performance depends on how the goals of the task are interpreted. A strict literal interpretation of the task is that perfect accuracy is required using as few selections as possible. By this measure, the crowd outperforms every individual because it is capable of perfect accuracy. Almost every participant in both frames fails to achieve this. The unweighted crowd reaches perfect accuracy with 15.8, 11.7, 42.7, and 43.2 selections for US present, US absent, Africa present, and Africa absent cases, respectively, while the confidence-weighted crowd needs 19.9, 11.1, 29.3. and 11.1 selections. It is clear that the weighted crowd outperforms the unweighted crowds in the relatively low-knowledge African countries condition.

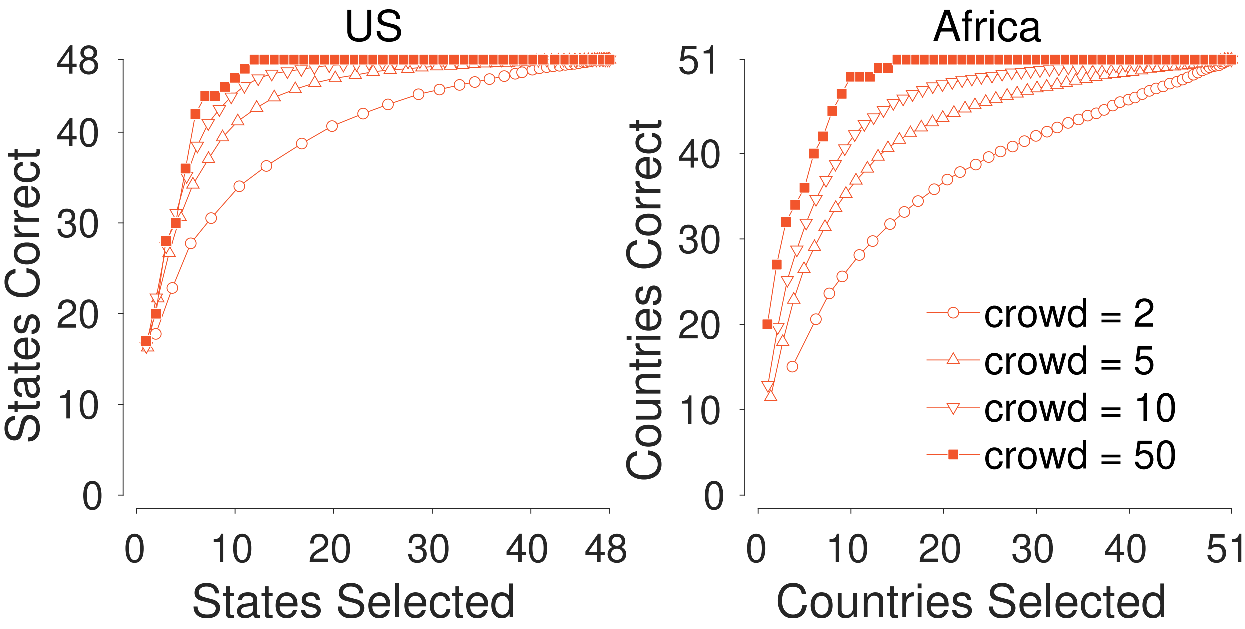

Figure 1.7: Performance of the crowd based on different numbers of individuals for both conditions. Each curve corresponds to the performance of the confidence-weighted crowd, including responses for both frames for crowds ranging from 2 to 50 individuals.

A less strict assessment of individual and crowd performance allows for less than perfect accuracy while still requiring relatively few selections. Visually, this corresponds to being at the top-left of the graphs shown in Figure 1.5. In the present frame of the US states condition, there are two participants whose performance is above and to the left of the crowd curve, and another three or four who are close. A similar result holds for the absent frame. In the African countries condition, there is one participant who meets this criterion in the present frame and none in the absent frame. A reasonable conclusion is that the crowd aggregate is superior to at least 90% of participants in the US states condition and essentially all participants in the African countries condition. For the vast majority of participants in all conditions and frames, the crowd's performance is both ordinally better and quantitatively much better.

1.5 The Wisdom of the Crowd Within

To examine the wisdom of the crowd within, we combined the selections made in the present and absent frames by the same participant for the same target. We also created crowd aggregate responses by combining the selections made by all of the participants in both frames. Figure 1.6 shows the results of these analyses. The blue dots correspond to individual participants, showing the average of the number of states or countries they selected over both framings, and the accuracy of their crowd-within aggregate. Accuracy is measured in terms of whether the correct state or country was selected in either the present or absent framing. The blue lines connect the aggregate individual performance to performance for just the present and absent frames separately (i.e., to the performance measures shown in Figures 1.3 and 1.5). These wisdom of the crowd within aggregates allow us to evaluate how perceptually similar participants treat the two structurally identical tasks as complementary responses would be exactly overlaid in Figure 1.6.

By its construction, the crowd-within aggregate always involves as many or more states or countries being selected as in the separate frames. Our interest is whether this increase significantly improves accuracy. Visually, this corresponds to crowd-within performance that shifts significantly upward without shifting far to the right. Figure 1.6 makes clear that, for most of the participants in both conditions, the crowd-within aggregate leads to an increase in accuracy. The mean increase in accuracy is 11.5 states and 17.9 countries. Much of this improvement comes from the absent frame selections broadening the selections to include the target as shown by the crowd-within aggregates moving diagonally toward the upper right in Figure 1.6. There are also cases in which two relatively narrow selections in the frame are combined to form an improved selection and where the crowd-within aggregates mainly shift upward with little movement to the right. For example, for the best performing individual in the US states condition, the crowd-within aggregate has perfect accuracy based on an average of 6.8 states being selected. This individual's crowd-within aggregate combined their present frame accuracy of 46 states, based on 6.0 selections, with their absent frame accuracy of 40 states, based on 5.0 selections. The crowd aggregation over both frames shown by the red curves continues to be well performed.

1.6 Crowd Size

Given the clear wisdom of the crowd effect, an interesting follow-up question is how many individuals are needed for effective crowd performance. Figure 1.7 shows the confidence-weighted crowd-within responses averaged over many subsets of 2, 5, or 10 randomly selected participants and the full crowd of 50 participants. The full crowd is the one considered in Figure 1.6, which uses all of the participant and frame information about each target. Both conditions show the same expected pattern of improved performance as the crowd size increases. There is an especially large improvement as the crowd increases from the smallest possible size of 2 to the still small size of 5. This pattern of initial quick improvement as the crowd size first grows followed by a long period of more gradual improvement is consistent with previous findings (Han & Budescu, 2019; Steegen et al., 2014; Vul & Pashler, 2008). For the US states condition, a crowd size of 10 is almost as well performed as the full crowd. For the African countries condition, the full crowd is clearly better performed than the smaller crowds. We interpret this result as showing that the more difficult African countries condition, about which participants had less knowledge, benefits more from incorporating more participants to capture the more sparsely distributed knowledge.

1.7 Discussion

We studied spatial knowledge in an experiment that asked participants to select regions on unlabeled tile maps to identify target US states or African countries. We asked for the knowledge to be expressed in two different ways, by framing the question in terms of identifying regions in which the target was present or eliminating regions from which the target was absent. Our first interest was in how people manage their uncertainty about the spatial location of the target, and whether this is affected by the different frames. Our second interest was whether wisdom of the crowd effects, including wisdom of the crowd within, are present for spatial knowledge.

We found that participants were consistently overconfident in their management of uncertainty, often to a very large degree. Many participants selected regions in the present frame that were too narrow and failed to include more than half of the targets. They were also overconfident in the absent frame, although to a lesser degree. The consistent pattern of overconfidence in both frames eliminates simple explanations in terms of minimizing effort and suggests that people are overconfident in their spatial knowledge. This sort of overconfidence is consistent with classic findings from the judgment and decision-making literature (Lichtenstein et al., 1982; Paese & Sniezek, 1991; Russo & Schoemaker, 1992; Welsh & Begg, 2018).

We also found strong wisdom of the crowd effects. Both unweighted and confidence-weighted aggregate crowd judgments outperformed the vast majority of individual participants. This was especially true for the more difficult African countries condition, suggesting most individuals have significant gaps in their knowledge but that collectively a crowd can perform well. At the individual level, we found that combining judgments from the same participants across both present and absent frames improved performance. A crowd aggregate that combined all participants and both frames achieved very good performance in both conditions. For the US states domain, a crowd of around 10 people proved enough to exhibit good performance, but the lower-knowledge African countries domain benefited from larger crowds.

Our results have implications both for understanding human cognition and practical applications. It is important to understand why people are overconfident in the regions they select, how robust this behavior is, and whether it can be mitigated. Future experiments should consider other spatial knowledge domains and other methods for expressing spatial knowledge, such as point estimates of locations or free-form selections of regions rather than discrete choices on tile maps. It is also important to understand how framing effects interact with the management of uncertainty. Our results suggest that the absent frame reduces overconfidence, but this could arise from the nature of the task design, and more robust replication is needed. In terms of practical applications, the demonstration of strong wisdom of the crowd effects holds promise for real-world problems like search and rescue (Breivik et al., 2013; Lin et al., 2013), military targeting (Council, 2013; Qing & Fang, 2021), and other problems where a spatial region needs to be identified based on human knowledge (e.g. Drew et al., 2013; Fu et al., 2017, 2020; Krupinski, 2010; Lin et al., 2014).

Finally, future work should apply cognitive modeling methods to understand people's behavior and potentially improve the wisdom of the crowd. This approach has proved fruitful in other cognitive domains including probability estimation, category learning, and sequential decision making (Danileiko & Lee, 2018; Lee & Danileiko, 2014; Thomas et al., 2021). Modeling how people select states and countries based on their knowledge should allow inferences about parameters that correspond to their uncertainty and decision-making strategies. A model-based approach to crowd aggregation may outperform the simple statistical methods on which our wisdom of the crowd results are based.

1.8 Publication Note

This chapter was previously published as Montgomery, L.E., & Lee, M.D. (2022). The wisdom of the crowd and framing effects in spatial knowledge. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the Annual Conference of the Cognitive*

Science Society, vol. 44. https://escholarship.org/uc/item/0h95m7m4.

Chapter 2

Where's Waldo, Ohio? Using Cognitive Models to Improve the Aggregation of Spatial Knowledge

Abstract

We apply cognitive modeling to improve the wisdom of the crowd in a spatial knowledge task. Participants provided point estimates for where 48 US cities are located and then, using the point estimate as a center point, chose a radius large enough that they believed the resulting circle was certain to contain the city's location. Simple and radius-weighted arithmetic averages of the individuals' point estimates produced more accurate group answers than the majority of individuals. These statistical aggregates, however, assume there are no differences in individual expertise nor in the difficulty of locating different cities. Accordingly, we develop a set of cognitive models to infer group estimates that make various assumptions about individual expertise and differences in city difficulty. The model-based estimates

generally outperform the statistical averages. The models are especially accurate if they allow for individual differences in expertise that can vary city by city. We replicate this finding by applying the same cognitive models to data reported by Mayer & Heck (2023) in which participants provided point estimates for the locations of European cities.

2.1 Introduction

The wisdom of the crowd is the idea that an aggregated judgment of a group of individuals is often more accurate than the judgments of the individuals in the group (Davis-Stober et al., 2014; Galton, 1907; Surowiecki, 2004). The basic premise is that crowd aggregation helps to minimize individual variability and error, while at the same time isolating the signal that contains the correct answer. The wisdom of the crowd has been broadly applied to tasks relating to general knowledge (Bennett et al., 2018; Lee & Danileiko, 2014; Prelec et al., 2017; Steyvers et al., 2009), forecasting or predictions (Butler et al., 2021; Himmelstein et al., 2023; Da & Huang, 2019; Klugman, 1947), and collaborative decision making (Knight, 1921; Lyon & Pacuit, 2013; Shaw, 1932). The elicited estimates from these tasks take various forms. Sometimes people give numerical answers, such as estimating when a historic event occurred (e.g., Herzog & Hertwig, 2009; Keck & Tang, 2020; Larrick et al., 2007). Sometimes people select between discrete options, such as choosing a country's capital city from a set of alternatives (e.g., Aydin et al., 2014; Simoiu et al., 2019). Sometimes people provide rankings, such as ordering a set of weights from lightest to heaviest (e.g., Gordon, 1924) or a list of cities from largest to smallest in terms of their population (e.g., Lee et al., 2014).

The wisdom of the crowd has also been applied to tasks that require spatial knowledge, such as locating cities on a map (Mayer & Heck, 2023) or selecting regions that include a state or country (Montgomery & Lee, 2022). Tasks like these involve making two-dimensional spatial estimates, emphasizing that the wisdom of the crowd is not restricted to scalar estimates or discrete choices. Spatial tasks also emphasize that expertise can be more complicated than a unidimensional measure of ability. It is reasonable to expect that people may be more expert at locating cities in geographic regions that they are familiar with, but there is also evidence that spatial estimates are affected by more abstract social and cultural categorical knowledge that varies across people (Friedman et al., 2002a,b, 2005, 2012).

One way to address the challenges of multidimensional behavior and structured expertise is to use cognitive models (see Lee, 2024, for an overview). The representational assumptions made by cognitive models provide a basis for aggregating multidimensional behavior, and the psychometric assumptions they make about individual differences provide a basis for inferring and up-weighting expertise. Cognitive models have been successfully used in wisdom of the crowd applications involving probability forecasts (Lee & Danileiko, 2014; Turner et al., 2014), rankings (Lee et al., 2014), category learning (Danileiko & Lee, 2018), competitive bidding (Lee et al., 2011b), combinatorial problem solving (Yi et al., 2012), and sequential decision tasks (Thomas et al., 2021). In all of these applications, the model-based approach forms crowd estimates without access to the ground truth or any other sort of normative feedback. The idea is that, as part of modeling people's observed behavior, the latent true values assumed to be generating the behavior can be inferred. These inferences constitute the model-based crowd estimates. Practically, because the model-based approach does not require any knowledge of the ground truth, it can be applied to real world problems involving spatial knowledge, such as search and rescue operations (Abi-Zeid & Frost, 2005; Lin & Goodrich, 2010; Wysokiński et al., 2014).

In this article, we use a cognitive modeling approach to improve the wisdom of the crowd aggregates for a spatial knowledge task similar to that developed by Mayer & Heck (2023). As for their task, we ask participants to provide point estimates of city locations. In addition, our task asks participants, starting at their point estimate, to extend a radius until they are certain that the resulting circle contains the true location of the city. We begin by

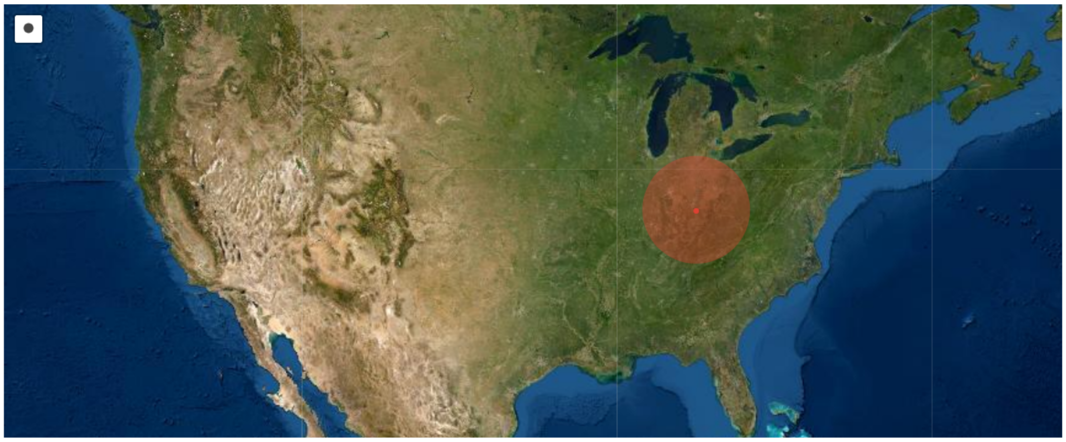

Figure 2.1: An example of a participant's response. Their point estimate of where the city is located is represented by the dark orange dot and their selected radius is represented by the larger orange circle surrounding it.

providing a description of our experiment and summarize the performance of individuals and statistical group aggregates. We then develop a series of cognitive models that make different assumptions about individual expertise, city difficulty, and whether or not to use the radius judgments. These models make many of the same assumptions as the Cultural Consensus Theory model developed by Mayer & Heck (2023), but also extend their modeling in key ways. We show that our model-based wisdom of the crowd estimates outperform the statistical wisdom of the crowd estimates, and that our model findings generalize to Mayer & Heck's (2023) data. We conclude with a discussion of theoretical implications of our findings for model-based wisdom of the crowd approaches, and the potential for applications.

2.2 Experimental Design

2.2.1 Experimental Interface

A screen shot of the experimental interface is shown in Figure 2.1. The interface displayed a contiguous map centered on the continental USA. There were no boundaries to distinguish the countries (the USA, Canada, and Mexico) or the 48 US states from each other. The interface was implemented using OpenStreetMap, a tiled web map with a geospatial data scheme similar to other popular interfaces such as Google Maps. The map was set to a fixed zoom level, and all methods of altering the zoom level, such as double-clicking or moving the mouse wheel, were disabled. These restrictions were intended to simplify the task and to standardize the correspondence between a participant's motor movement and their level of assumed uncertainty in specifying a radius.

2.2.2 Participants

A total of 50 participants were recruited on Prolific (`www.prolific.co`) to complete the task. The youngest participant was 19, the oldest participant was 61, and the median age was 32. All participants were current US residents who had attended high school in the USA. They were each asked which US states they were familiar with, which was operationalized as the states that they had lived in previously or visited frequently. All participants were familiar with at least one state, and 27 participants reported being familiar with more than one state. The maximum number of familiar states reported was 19.

2.2.3 Procedure

Participants were asked to estimate where a set of 48 cities, containing the most populous city in each of the contiguous US states, were located. They began the task by watching a 3-min video demonstrating how to select a point on the map and indicate a radius around it. The video emphasized that participants should select the initial point that represented their "best estimate" of each city's location before dragging their mouse outward to the desired radius, stopping when they were certain that the city's true location was within the area of the circle. Participants were specifically told to "first make your best guess and expand your radius of uncertainty from there," with the goal of "stopping when you're certain the location is within the area of your circle." The full instructions can be found in the supplementary material. Radius judgments were allowed to go beyond land borders and encompass surrounding bodies of water. Figure 2.1 shows an example of a participant's response. The point estimate for the city's location is shown by the dark orange dot, and the judged radius generates the larger surrounding orange circle.

At the start of the task, all participants were given a practice trial in which they were asked to locate San Francisco, California. Responses in this practice trial were not recorded. Participants then completed the main task in which they provided a point estimate and radius judgment for the 48 cities. The order of cities was randomized for each participant. On each trial, participants could redo their point estimate and radius judgment as many times as they liked before moving on to the next city. Only their most recent selection for each city was recorded, and participants were not allowed to return to an earlier city. There was no time constraint on individual trials, but the entire task had to be completed within the allotted time on Prolific, which was 87 minutes. On average, participants took 23.5 minutes to complete the task and answer the demographic questions after having watched the instructional video. Participants were not provided with any feedback on either the practice trial or the main trials. We did not exclude any responses.

We normalized the latitude and longitude spatial estimates provided by the experimental software to be consistent with the physical dimensions of the map in the interface, which was approximately 2.44 times wider than it was tall. This means that x-axis and y-axis spatial locations on the normalized scale took values between (0, 2.44) and (0, 1), respectively. The experimental software provided radius judgments in terms of miles, which we converted into degrees of latitude in the North direction to map them to the normalized scale. For both the point estimates and radius judgments, we ignored the Earth's curvature.

2.3 Behavioral Analyses

2.3.1 Participant Performance

Given the true locations of the 48 cities and the point estimates and the radius judgments provided by participants, we measured participant performance two different ways. The first *mean error* measure considered how far away point estimates were from true locations, which we calculated as the mean Euclidean distance on the normalized scale. The second *accuracy* measure considered the proportion of circles around the point estimate that contained the true location. Over all participants, the mean error was 0.13, the mean radius was 0.17, and the resulting circles were correct 64% of the time. The two measures of performance—mean error and accuracy—had a correlation of $r = -0.54$, meaning that participants with better point estimates tended also to include the target cities in their circles. The correlation between mean error and the mean radius judgment was $r = 0.35$, meaning that participants with worse point estimates tended to express more uncertainty.

As examples of individual participant behavior, Figure 2.2 shows the performance of a relatively well-performed and a relatively poorly-performed participant. Each city's true location is shown as a black square. A black line connects the true location to the point estimate of the participant. The circles that surround the point estimates show the radius judgment of the participant, and are color-coded so that an accurate response is blue and an inaccurate response is red. The well-performed participant had a mean error of 0.033, provided an average radius of 0.075, and their circles contained the true location 83% of the time. The poorly-performed participant had a mean error of 0.18, provided an average radius of 0.14, and their circles contained the true location only 33% of the time.

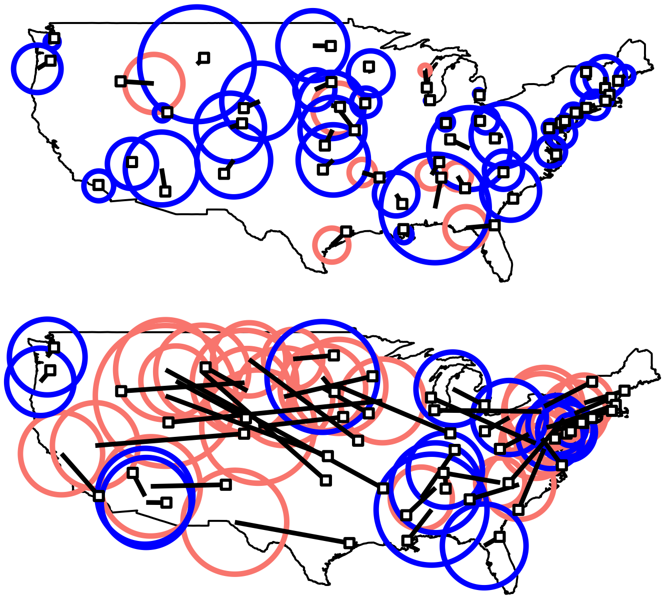

Figure 2.2: The true locations of the 48 city locations compared with the estimated locations for a well-performed (top) and poorly-performed (bottom) participant. The true locations of the cities are shown by squares, and the error is shown by the line connecting the true location to the participant's point estimates. The circles generated by the point and radius estimate are shown in blue if they contain the true location and in red if they do not.

2.3.2 Crowd Performance

We used the arithmetic mean and a weighted arithmetic mean as *statistical* wisdom of the crowd estimates. The simple wisdom of the crowd estimate is the unweighted average of the individual participants' estimates: for city j, it is $\frac{1}{n}\sum_{n=1}^{i}\boldsymbol{y}_{ij}$, where $\boldsymbol{y}_{ij}$ is the point estimate of participant i for city j. The weighted wisdom of the crowd estimate is a weighted average of the individual participant estimates according to the area of the circle they provided: for city j, it is $\frac{1}{n}\sum_{n=1}^{i}\frac{1}{r_{ij}^2}\boldsymbol{y}_{ij}$, where r_{ij} is the radius judgment of participant i for city j. The weighted wisdom of the crowd estimate puts more weight on the estimates of individuals who provided a smaller radius judgment and thus identified a smaller possible area in which the city could be located.

Figure 2.3 provides four examples of individual estimates producing crowd aggregate estimates. These are Jacksonville in coral, Seattle in teal, Houston in lilac, and Boise in green.

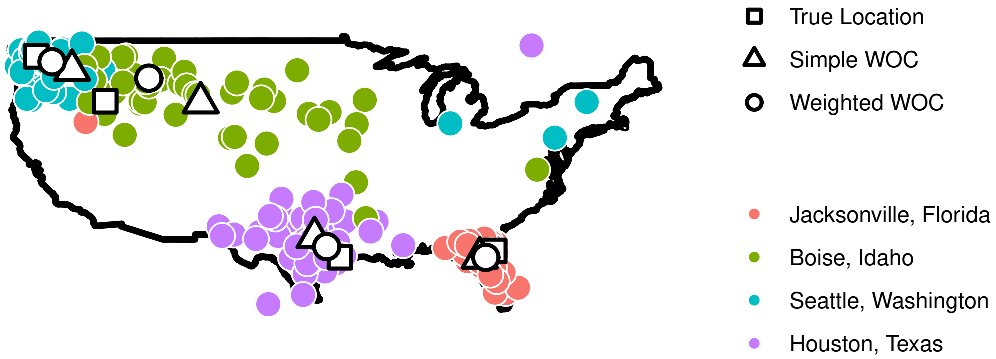

Figure 2.3: The 50 participants' estimates for four cities: Jacksonville (coral), Boise (green), Seattle (teal), and Houston (lilac). The city's true location is shown as a square, the simple wisdom of the crowd estimate is shown as a triangle, and the weighted wisdom of the crowd estimate is shown as a circle.

For all four cities, the true target location is shown as a square, and the simple and weighted crowd estimates are shown as triangles and circles, respectively. The crowd estimates are generally closer to the true location of the city than most of the individual estimates. In addition, the weighted wisdom of the crowd estimates tend to be closer to the target location than the simple wisdom of the crowd estimates.

Comparing the four cities, Figure 2.3 demonstrates clear differences in how difficult different cities were to locate. Jacksonville had a mean error across all participants of 0.079 and was the city most often correctly contained in participants' circles, with 86% accuracy. Seattle had a mean error of 0.13, with 78% accuracy. Houston was slightly more difficult for participants to locate. The mean error was 0.14, and accuracy was 68%. Boise was one of the most difficult cities to locate with a mean error of 0.25 and only 28% accuracy.

The examples in Figure 2.3 provide the insight that cities may have different inherent difficulties, not just in relation to each other, but also in terms of differences in locating the correct longitude versus latitude. Seattle appears to be easier for participants to locate than Boise, and the uncertainty for Seattle seems to be approximately circular. Boise, in addition to being more difficult, appears to be more difficult along its longitude than its latitude.

This unequal difficulty results in the uncertainty for Boise across participants being elliptical in shape. Jacksonville, in contrast, looks to be more difficult along its latitude than longitude, likely because participants use the constraining geographic information provided by the coastline of the peninsula.

2.4 Cognitive Models for Aggregating Estimates

A cognitive model of participant behavior in our task needs to consider both the point estimates and radius judgments that the participants made. We describe the model of behavior in terms of these two parts.

2.4.1 Model of Point Estimates

Our approach to modeling the point estimate uses several key features of the cognitive model developed by Mayer & Heck (2023). We adopt the same basic assumption that the point estimate $\boldsymbol{y}_{ij}$ is sampled from a bivariate Gaussian distribution centered on the latent true location of the city, with a potentially tilted elliptical shape that represents the uncertainty the participant has about the location. Formally, our model assumes that

$$\boldsymbol{y}_{ij} \sim \text{bivariate Gaussian}\big(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{ij}\big), \tag{2.1}$$

where $\boldsymbol{\mu}_j$ is the unknown latent location of city j, and the uncertainty about its location is captured by the covariance matrix $\boldsymbol{\Sigma}_{ij}$. The latent true location has both a longitude μ_{j1}

and latitude μ_{j2} with prior distributions that are uniform over the normalized scale:

$$\begin{aligned}\mu_{j1} &\sim \text{uniform}\big(0, 2.44\big) && (2.2)\\ \mu_{j2} &\sim \text{uniform}\big(0, 1\big). && (2.3)\end{aligned}$$

It is the inferences made by the model about these parameters from people's data that corresponds to the model-based wisdom of the crowd aggregate.

The covariance matrix $\mathbf{\Sigma}_{ij}$ in Equation 2.1 is specified as

$$\boldsymbol{\Sigma_{ij}} = \begin{bmatrix} \lambda_{j1}^2 + \sigma_i^2 + \beta_{ij}^2 & \rho_j \sqrt{\lambda_{j1}^2 + \sigma_i^2 + \beta_{ij}^2}\sqrt{\lambda_{j2}^2 + \sigma_i^2 + \beta_{ij}^2} \\ \rho_j \sqrt{\lambda_{j1}^2 + \sigma_i^2 + \beta_{ij}^2}\sqrt{\lambda_{j2}^2 + \sigma_i^2 + \beta_{ij}^2} & \lambda_{j2}^2 + \sigma_i^2 + \beta_{ij}^2 \end{bmatrix}, \quad (2.4)$$

which incorporates the overall expertise of individual i, σ_i, the city-specific expertise of individual i for city j, β_{ij}, the difficulty of city j with respect to its longitude λ_{j1} and latitude λ_{j2}, and a correlation ρ_j. There are two expertise components included in the covariance matrix: one for the individual's overall expertise and one for their city-specific expertise. The individual's overall expertise σ_i is a measure of the average uncertainty they have across all cities. Smaller values of σ_i correspond to reduced uncertainty and greater expertise. The city-specific expertise β_{ij} provides an offset to the average uncertainty for each city. It is modeled hierarchically with a mean of zero and variance ω_i^2:

$$\beta_{ij} \sim \text{Gaussian}_{+}\big(0, \frac{1}{\omega_i^2}\big). \quad (2.5)$$

The model developed by Mayer & Heck (2023) similarly included individual expertise and city difficulty components, but our introduction of a city-by-expertise component is new. Restricting the model to just individual expertise corresponds to assuming that individuals can be more or less expert than each other, but that an individual is equally expert for all cities. Our motivation for including individual-by-city expertise is to allow individuals to have some city-specific knowledge. The value of β_{ij} increases or decreases the average expertise of individual i in the specific context of city j. Larger values of ω_i mean that an individual's expertise differs more from city to city. We use diffuse priors on the individual expertise and the variability in individual-by-city expertise parameters:

$$\begin{aligned} \sigma_i &\sim \text{uniform}\big(0,1\big) && (2.6)\\ \omega_i &\sim \text{uniform}\big(0,1\big). && (2.7) \end{aligned}$$

We divide a city's difficulty into a longitude difficulty λ_{j1} and latitude difficulty λ_{j2}. Separating a city's difficulty into these two parts is based on the intuition, made clear in Figure 2.3, that some cities are more difficult to locate along one of these dimensions. We assume that these difficulties are hierarchically distributed, using diffuse priors:

$$\begin{aligned} \lambda_{j1} &\sim \text{Gaussian}_{+}\big(\mu_{\lambda_1}, 1/\sigma^2_{\lambda_1}\big) && (2.8)\\ \lambda_{j2} &\sim \text{Gaussian}_{+}\big(\mu_{\lambda_2}, 1/\sigma^2_{\lambda_2}\big) && (2.9)\\ \mu_{\lambda_1}, \mu_{\lambda_2} &\sim \text{uniform}\big(0,2\big) && (2.10)\\ \sigma_{\lambda_1}, \sigma_{\lambda_2} &\sim \text{uniform}\big(0,1\big). && (2.11) \end{aligned}$$

The correlation ρ_j completes the statistical representation of an uncertainty ellipse that can vary in orientation, and is also given a diffuse prior

$$\rho_j \sim \text{uniform}\big(-1, 1\big). \tag{2.12}$$

2.4.2 Model of Radius Judgments

Mayer & Heck (2023) did not collect or attempt to model radius judgments, so this part of our model is entirely new. The key assumption we make for the radius y^r_{ij} is that it depends both on the uncertainty ellipse and how a participant manages that uncertainty to produce a circle that expresses their confidence. The variances of the ellipse are provided by the diagonal elements of the covariance matrix in Equation 2.4. Given that the experimental task constrained participants to use circles, it seems reasonable to assume radius judgments were based on the largest standard deviation $\sqrt{\max(\boldsymbol{\lambda}_j)^2 + \sigma_i^2 + \beta_{ij}^2}$. We then assume that there are individual differences in how participants manage their uncertainty using a scale parameter α_i for individual i. Formally, our model assumes that the radius judgment is

$$y^r_{ij} \sim \text{Gaussian}\big(\alpha_i \sqrt{\max(\boldsymbol{\lambda}_j)^2 + \sigma_i^2 + \beta_{ij}^2}, 1/\tau^2\big). \tag{2.13}$$

Thus, the scale parameter effectively corresponds to how many standard deviations, in the direction of maximum uncertainty, participants use to determine their radius judgments. The parameter τ measures the precision with which participants produce intended radius judgments in the experimental interface. It is a measure of motor movement error and other sources of noise, and is assumed to be common to all individuals on all trials. Both the uncertainty scaling α_i and response noise τ parameters are given diffuse priors:

$$\begin{aligned} \alpha_i &\sim \text{uniform}\big(0, \sqrt{2.44^2 + 1^2}\big) && (2.14) \\ \tau &\sim \text{uniform}\big(0, 1\big). && (2.15) \end{aligned}$$

2.4.3 Model Identifiability

The full cognitive model defined by Equations 2.1–2.15 defines a joint model of the point estimate and radius judgments. To test whether the model is identifiable, especially given the introduction of flexibility by allowing for individual-by-city expertise, we conducted a simulation study. We created 50 artificial participants using the posterior means found by applying the model to the participants in our task. The motivation was to make sure the artificial participants had a realistic range of parameter values. We then simulated 50 experiments in which the model was used to generate artificial point estimates and radius judgments for each participant and city. Finally, we applied the model to make inferences from the simulated data. The inferences approximated the known generating values for all parameters, both in the aggregate across experiments and (especially) by averaging over experiments. The code, simulated data, and results associated with this parameter recovery study can be found in the supplementary information.

We conclude from the successful parameter recovery that the model is identifiable. We speculate that there are two main reasons for this. One is that most of the model's key parameters—individual expertise, individual-by-city expertise, and city difficulty—play a role in making predictions about both the point estimates and radius judgments. This makes the model constrained in terms of its joint prediction of the two different components of the behavioral data. The second likely basis for identifiability lies in the constraints inherent in two-dimensional spatial judgments coming from the metric axioms that define distances in the space.

2.4.4 Model Variants

The full model has three important features. The first is that expertise varies not only by individual σ_i, but also by individual and city β_{ij}. The second is that each city has its own difficulty $\boldsymbol{\lambda}_j$ that is specified in terms of separate longitude and latitude difficulties. The third is that both the individual's point estimate and their radius judgment are included. Simplified models can be constructed by changing one or more of these features, and serve to test whether or not the various features of the model contribute to good wisdom of the crowd aggregation.

For expertise, we consider two simpler assumptions than the full model: that there are no individual differences and all individuals have the same expertise σ or that there are individual differences in expertise σ_i but individuals do not have a city-specific expertise. To switch between the total of three different assumptions about expertise requires changing Equations 2.4 and 2.13 to use either σ, σ_i, or σ_i and β_{ij}. For the assumption of no individual differences in expertise, it is also necessary to remove Equations 2.5 and 2.7 and replace σ_i in Equation 2.6 with σ. The assumption of no individual-by-city expertise requires removing Equations 2.5 and 2.7.

For city difficulty, we consider the alternative assumption that city difficulty is still different in terms of longitude and latitude, but that these difficulties no longer vary by city. Instead, all cities share the same longitude difficulty λ_1 and latitude difficulty λ_2. To make these assumptions, it is necessary to remove Equations 2.10–2.11 and adjust Equations 2.8–2.9. Specifically, λ_{j1} and λ_{j2} in Equations 2.8–2.9 become $\lambda_1, \lambda_2 \sim \text{uniform}(0, 2)$.

For radius judgments, we consider the possibility of only modeling the point estimates. This requires removing Equations 2.13, 2.14, and 2.15. This specific form of the model can be applied to data involving only point estimates, such as those collected by Mayer & Heck (2023).

2.4.5 Model Implementation

Exhaustively combining the three assumptions about expertise, the two assumptions about city difficulty, and whether or not the radius judgments are included produces 12 different models. We implemented all of these models as graphical models in JAGS (Plummer, 2003) to allow for fully Bayesian inference based on computational sampling approximation to the joint posterior (Lee & Wagenmakers, 2014). Our results are based on six independent chains each with 5000 samples, a burn-in of 1000 samples, and thinning the chains by retaining one in every 4 samples. We evaluated the chains for convergence according to the standard $\hat{R}$ (Brooks & Gelman, 1998) measure. JAGS and R code for the modeling analysis is available in the supplementary information.

2.5 Results

2.5.1 Performance Results

Each of the 12 models makes a prediction about where each of the 48 cities are located. These predictions are the inferences for the latent true location $\boldsymbol{\mu}_j$ of the cities, and take the form of a posterior distribution over the two-dimensional map. The posterior distribution quantifies how likely it is that every location on the map is the true location of a city, based on the observed estimates people made for the city, and the cognitive modeling assumptions

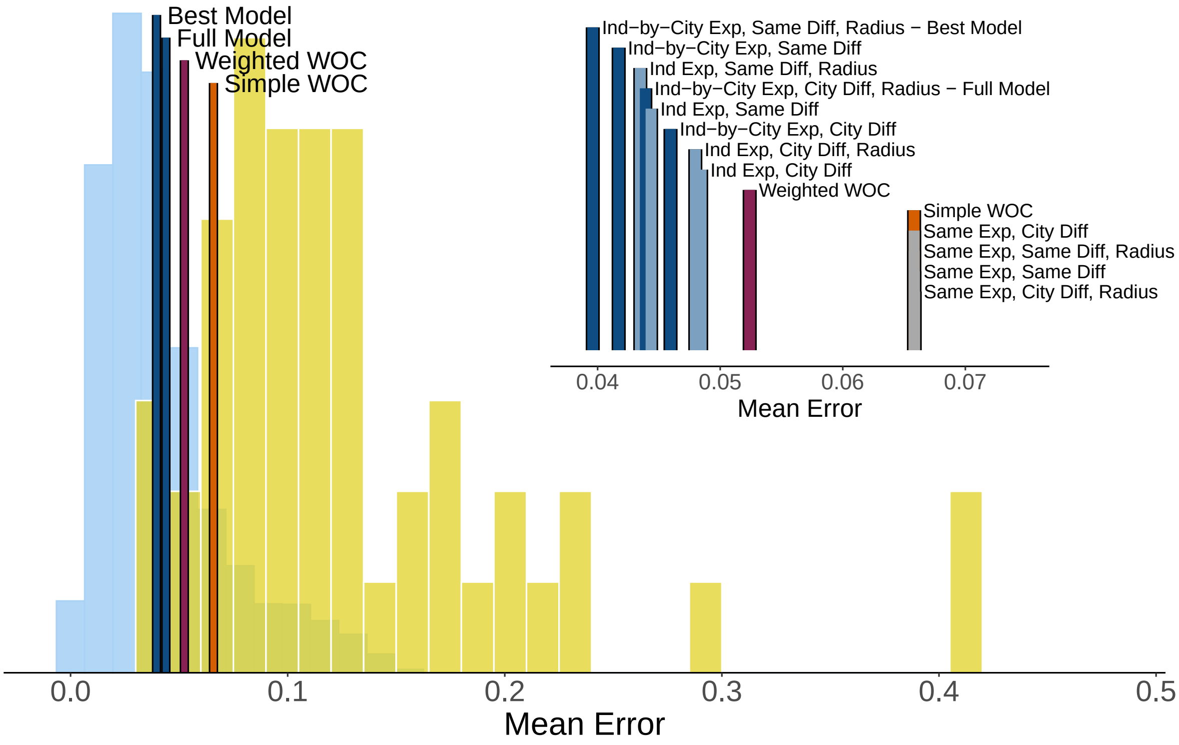

Figure 2.4: The main panel shows the distribution of individual mean error in yellow and the mean error of statistical and model-based crowd aggregates by vertical lines. The posterior distribution for the best-performing model's mean error is shown in blue. The vertical bars in the inset panel provide a magnified view of the performance of model-based and statistical estimates, with color coding to indicate the assumption each model makes about expertise.

about how they produced those estimates. As emphasized above, these inferences are made without access to the ground truth. Once the inferences have been made, however, it is possible to measure their performance by comparing them to the true locations of the cities. The posterior distribution can be used to construct a posterior distribution of the error of the model. A convenient simpler point estimate measure of error is the distance between the posterior mean and the true city location.

The main panel of Figure 2.4 shows how the wisdom of the crowd estimates for the various models compare to individual performance and the performance of statistical aggregates. The mean error of the individuals are shown as a yellow histogram. The mean error of the statistical wisdom of the crowd estimates, and two of the model-based estimates, are shown as vertical lines. The best-performing model assumes that there is individual expertise that varies across cities and includes the radius judgments, and has point estimates that are on average 0.040 from the true locations of the cities. This model's full posterior distribution of mean error is shown in light blue. There is evidence of a wisdom of the crowd effect, because all of the statistical and model-based crowd estimates outperform the majority of individuals in the crowd. There is further evidence that model-based estimates outperform statistical estimates in aggregating individual knowledge.

The inset bar plot in Figure 2.4 compares the different wisdom of the crowd estimates to each other, focusing on the restricted range of mean error in which they all lie. The two statistical wisdom of the crowd estimates are the simple wisdom of the crowd estimate in orange and weighted wisdom of the crowd estimate in maroon. The other 12 lines correspond to the 12 cognitive models. The lines are labeled according to how they incorporate expertise, city difficulty, and the radius judgments. The line color corresponds to how the model incorporates expertise: gray lines indicate that the model assumed no individual differences in expertise, light blue lines indicate that the model assumed individual differences, and dark blue lines indicate that the model assumed individual differences that vary by city. The models that allow for individual expertise outperform the models that assume expertise is constant across participants, and generally, the models that include the individual-by-city expertise perform better than models with just individual expertise. Further interpretation of these results may not generalize beyond this data set, but we think that the pattern of results suggests that assumptions about expertise affect the performance of crowd estimates. Our results also suggest that there may be some trade-off between including the radius judgments and assuming individual-by-city expertise, so that models with either tend to do better than models without.

2.5.2 Parameter Results

Our main focus in evaluating the cognitive models is on predictive accuracy, but a different way to use the models is as measurement models. The parameters correspond to meaningful psychological properties like expertise, uncertainty management, and city difficulty. Figure 2.5 shows the inferences about key parameters from the full model for all participants, and how they relate to basic behavioral measures. In all of the panels, the model parameters are represented by their posterior mean and their 95% credible interval. The correlations between the model parameters and their corresponding behavioral measures are provided in the bottom-right-hand corner of each panel.

The two panels in the top row of Figure 2.5 focus on the expertise and uncertainty management of individuals. The top-left panel compares the model's inferences of individual expertise σ_i to the behavioral measure of performance provided by the mean error of an individual's point estimates. Individuals with smaller errors had smaller σ_i, consistent with greater expertise. We emphasize again that the model was not provided information about the cities' true locations, so the correlation of σ_i with performance shows that the model is genuinely able to predict the relative expertise of individuals. The top-right panel compares model inferences about an individual's management of uncertainty α_i with their average radius. Individuals inferred to express more of their latent uncertainty gave larger average radius judgments.

The two panels in the bottom row focus on the cities instead of individuals. The bottom-left panel compares the model's inferred maximum city difficulty across longitude and latitude, $\max \boldsymbol{\lambda}_j$, to a behavioral measure of city accuracy. This measure was calculated in the same manner as individual accuracy. Instead of measuring how far a particular individual's estimates were from the true locations, we measured how far on average the estimates for a particular city across individuals were from the city's true location. Cities that were inferred

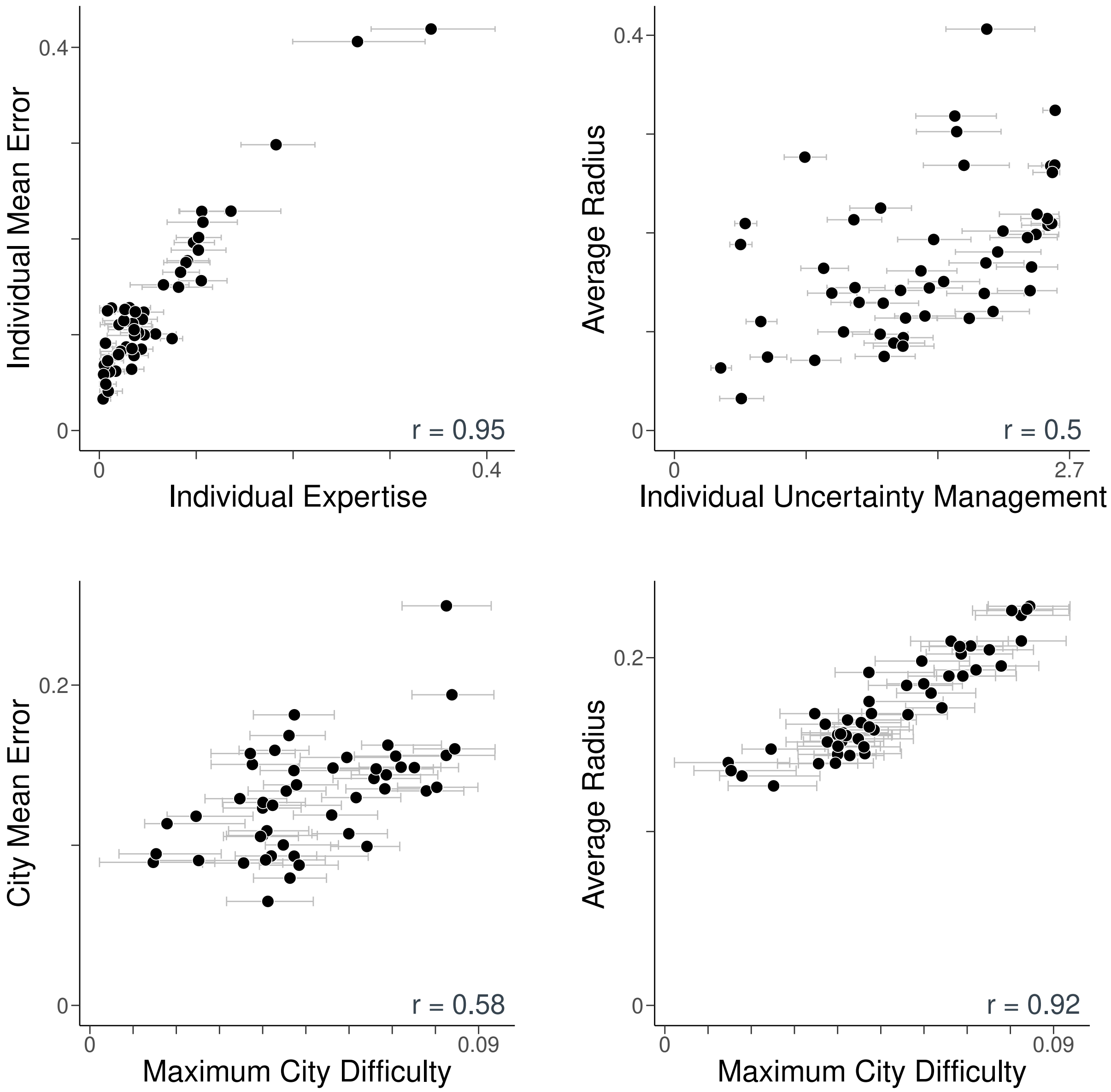

Figure 2.5: The relationship between parameter values and behavioral measures of individual differences in terms of expertise and uncertainty, and city differences in terms of difficulty. See main text for details.

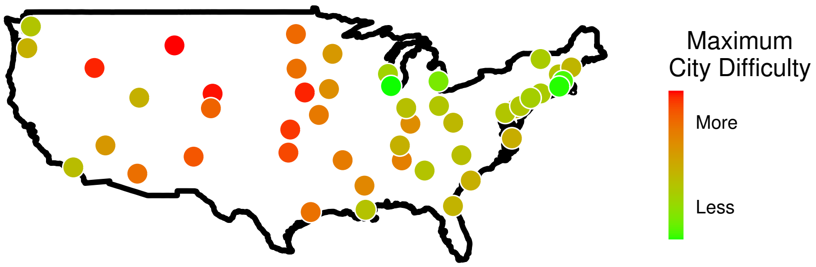

Figure 2.6: A visualization of inferred maximum city difficulty. Circles are located on the true locations of the cities. Cities inferred to be less difficult are in bright green, and cities inferred to be more difficult are in bright red.

to be easier to locate had smaller mean errors, while cities that were inferred to be harder to locate had larger mean errors. Once again, because the model is not provided with ground truth information, these are predictions about relative city difficulty. The bottom-right panel compares the inferred city difficulty to the average radius size for that city. Cities that were more difficult had larger average radius sizes.

The results in Figure 2.5 show that the key model parameters of expertise, uncertainty management, and city difficulty correlate well with conceptually related behavioral measures. Figure 2.6 demonstrates one way that these parameters can be used for interpretation. It shows the inferred difficulties of the cities, ranging from the most difficult in bright red to the easiest in bright green. The cities on the east and west coasts were generally inferred to be less difficult than those that were more centrally located.

Figure 2.5 does not include a comparison of the city-specific expertise β_{ij} with a behavioral measure. Of the experimental data we collected, the most likely candidate is the individual's familiarity with different states. Using the self-reported familiarity information, we compared the distribution of individual-by-city expertise for cities that were in familiar states with cities that were in unfamiliar states. These distributions were extremely similar, and had a mean difference of only 0.005. Accordingly, it seems that individual-by-city expertise, as incorporated in our model, is sensitive to some other information than self-reported familiarity.

2.5.3 Application to Mayer & Heck (2023)

To evaluate the replicability and generalizability of our findings, we applied the same cognitive models to the data set collected by Mayer & Heck (2023). Mayer & Heck (2023) had 228 participants provide point estimates for 57 European cities on seven different maps of Austria and Switzerland, France, Italy, Spain and Portugal, the UK, Eastern Europe, and Germany. We followed Mayer & Heck (2023) in excluding participants who gave point estimates that were outside the countries of interest for more than 10% of the cities. Participants were not asked to provide radius judgments, so we only applied the models using point estimates. We tested the six models that exhaustively combined the three assumptions about expertise and the two assumptions about city difficulty.

We also compared our model's performance with the model developed by Mayer & Heck (2023). Their model was inspired by the Cultural Consensus Theory model for two-dimensional judgments known as CCT-2D (Anders et al., 2014; Romney et al., 1986). Cultural Consensus Theory was developed in cultural anthropology as a model of crowd consensus in the absence of ground truths. A simple example is a society agreeing that the number 13 is unlucky. We think that Mayer & Heck's (2023) application of Cultural Consensus Theory to the location of cities, which have objective ground truths, reduces to a model-based wisdom of the crowd approach. Because of its CCT-2D foundations, there are a few differences in the details of Mayer & Heck's (2023) cognitive model when compared to ours, related to the scales on which parameters are defined and the priors they are subsequently given. However, at its heart, their model assumes that individuals possess some cultural competence, which we think of as synonymous with individual expertise in this context, and that items have variable difficulty in two different dimensions. We think that this makes the CCT-2D model

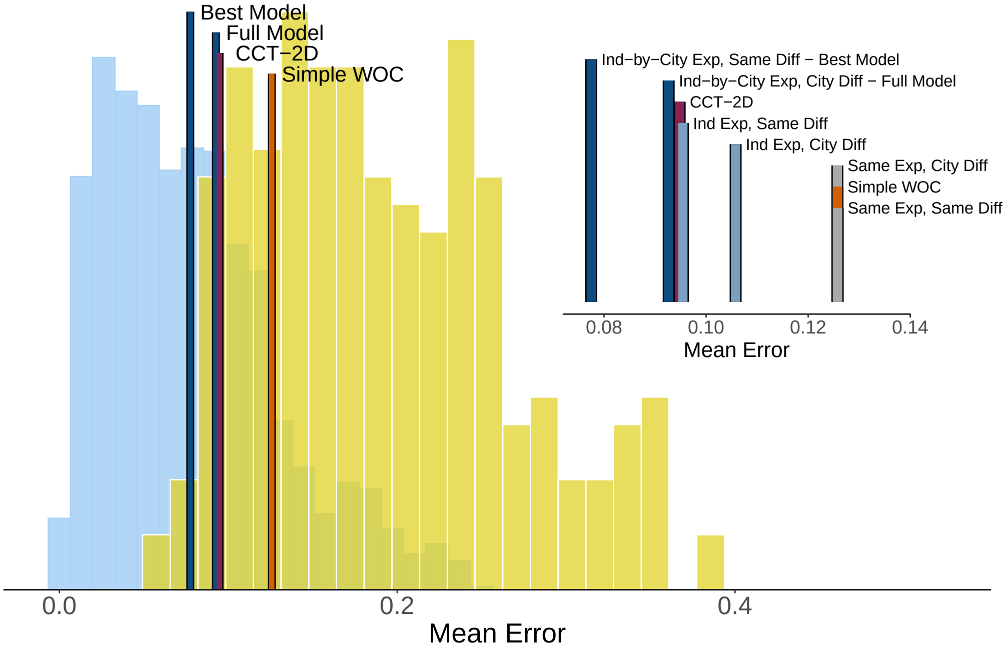

Figure 2.7: Analysis of Mayer & Heck's (2023) data. The main panel shows the distribution of individual mean error in yellow and the mean error of statistical and model-based estimates by vertical lines. The posterior distribution for the best-performing model's mean error is shown in blue. The vertical lines in the inset panel provide a magnified view of the mean error of model-based and statistical estimates, with color coding to indicate the assumption each model makes about expertise.

conceptually the same as our model that assumes individual expertise and allows for city difficulty, but does not allow for individual-by-city expertise or incorporate radius judgments.

The performance of our models and the Mayer & Heck (2023) model are shown in Figure 2.7. The simple wisdom of the crowd estimate again outperforms the majority of individuals demonstrating that there is a wisdom of the crowd effect. The best-performing model allows for individual-by-city expertise, but assumes cities have equal difficulty. Its inferred city location point estimates have a mean error of 0.077 from the true locations of the cities. The second-best model additionally allows for variable city difficulty. Overall, the model-based wisdom of the crowd estimates improve as the expertise assumption changes from having no individual differences to having individual differences and then finally to individual differences that also vary by city. The models that assume no individual differences in expertise perform very similarly to the simple statistical wisdom of the crowd estimate.

These modeling results replicate the key finding from our experiment by showing improved performance by allowing individual-by-city expertise. For both data sets, it generally appears that the models allowing for individual-by-city expertise but not variable city difficulty perform the best. The application to Mayer & Heck's (2023) data also underscores the point that our modeling approach can infer expertise based only on the point estimates of city locations. Finally, it is interesting to note that the Mayer & Heck (2023) model performed slightly better than our model that made the same psychological assumptions, presumably due to their different priors.

2.6 Discussion

We found a wisdom of the crowd effect in the spatial estimation problem of locating cities. Statistical aggregates of people's estimates outperformed most individual estimates. We also found that cognitive models can outperform both the simple and weighted statistical aggregations. Model-based estimates improved the wisdom of the crowd estimates primarily because they allowed for differences in individual expertise. We also found a consistent but smaller improvement associated with allowing for individual-by-city expertise in addition to individual expertise.

Most previous cognitive models used to find the wisdom of the crowd have assumed that expertise is a stable property of the individual across all of the items in the domain being judged (e.g. Lee & Danileiko, 2014; Lee et al., 2012, 2014; Mayer & Heck, 2023). Our findings suggest this assumption could be too simple. Conceptually, allowing people to have different levels of expertise for different items changes the emphasis on how the wisdom of the crowd is achieved. For the wisdom of the crowd effect, Lee (2024) distinguishes between a *signal and* *noise* mechanism that relies on aggregating judgments to amplify common signal and cancel noise, and a *jigsaw puzzle* mechanism that relies on diversity in knowledge so that different people provide accurate answers to different subsets of a problem. The use of individual-by-city expertise recognizes this diversity and allows the weight of an individual's estimate to be different for different cities. We do not yet, however, have a good account of how and why expertise varies across items. The basic hypothesis that for city locations people's expertise is related to their self-reported familiarity with those cities was not supported by our data.

Expertise has been explored before in the wisdom of the crowd literature. Others have investigated how smaller select crowds of experts can be more accurate than larger ones (Mannes et al., 2014; Olsson & Loveday, 2015) and found ways of identifying those with more relative expertise within the crowd (Budescu & Chen, 2014; Goldstein et al., 2014). Smaller select crowd performance has also compared different crowd compositions, like those of novices or experts (Fiechter & Kornell, 2021), and into the specific conditions that must be met for smaller select crowds to be more accurate (Davis-Stober et al., 2014, 2015). Most of this research, however, has also viewed expertise as a relatively stable personal trait. Future work should explore structured context-dependent accounts of expertise. Our modeling allowed for individual-by-city expertise, but lacked a theory to understand how and why expertise varied. One possible approach is to use hierarchical representations of expertise in terms of general and specific abilities, of the type that form the foundation of psychometric studies of cognitive abilities (Deary, 2020; McGrew, 2009). There are also structural accounts of expertise within specific domains that could be especially useful in the wisdom of the crowd context (e.g., Schvaneveldt et al., 1985).

Future work could also explore other sorts of spatial estimation tasks. For example, our task restricted people to providing circles to represent their spatial knowledge. This simplifies the task and the analysis, but it would be interesting to allow people to draw free-form shapes that could better express their knowledge. We also provided simple instructions of extending a circle until people were confident they had included the city. It would be possible to be more precise, and ask people (for example) to be 95% certain, although findings on the calibration of probability judgments suggest that people may not be able to do this well, since they are often overconfident (Hora, 2004; Keren, 1991; Lichtenstein et al., 1977; Ronis & Yates, 1987; Wallsten et al., 1993). Bigger variations on the basic task are also possible. For example, Montgomery & Lee (2022) asked participants to select a region on a map, instead of a point estimate and radius. The task also required manipulating the way the spatial knowledge question was framed, by asking participants either to select a region that included the target or select all the regions that did not include the target. Thus, for example, participants were asked to select as few US states as as possible an unlabeled map so that Ohio was included in the selection, or as many states as possible without including Ohio. A model-based wisdom of the crowd approach thus would need to understand how the question framing affected the participant's management of their uncertainty about Ohio's location. The extra complexity required in modeling people's behavior, however, has the benefit of allowing multiple estimates to be collected from the same individual, consistent with the wisdom-of-the-crowds-within effect (Herzog & Hertwig, 2014; Vul & Pashler, 2008).

Spatial knowledge provides an interesting application of the cognitive modeling approach to the wisdom of the crowd. Our modeling analysis suggests that expertise is best treated as multidimensional, and demands a representation that allows for people's expertise to vary across the spatial domain. This finding emphasizes that the wisdom of the crowd is not just a statistical consequence of reducing noise by sampling many people, but also a psychological consequence of incorporating enough people in a crowd to capture a diverse range of knowledge. It seems likely that cognitive modeling approaches to the wisdom of the crowd in other settings will benefit from allowing this diversity in their representations of individual differences.

2.7 Publication Note

This chapter was previously published as Montgomery, L.E., Baldini, C.M., Vandekerckhove, J., & Lee, M.D. (2024). Where's Waldo, Ohio? Using cognitive models to improve the aggregation of spatial knowledge. *Computational Brain & Behavior*, *7*(2), 242–254. `https://doi.org/10.1007/s42113-024-00200-0`

Chapter 3

The Wisdom of the Crowd with Partial Rankings: A Bayesian Approach Implementing the Thurstone Model in JAGS

Abstract

We develop a Bayesian method for aggregating partial ranking data using the Thurstone model. Our implementation is a JAGS graphical model that allows each individual to rank any subset of items, and provides an inference about the latent true ranking of the items and the relative expertise of each individual. We demonstrate the method by analyzing data from new experiments that collected partial ranking data. In one experiment, participants were assigned subsets of items to rank; in the other experiment, participants could choose how many and which items they ranked. We show that our method works effectively for

both sorts of partial ranking in applications to US city populations and the chronology of US presidents. We discuss the potential of the method for studying the wisdom of the crowd and other research problems that require aggregating incomplete or partial rankings.

3.1 Introduction

The wisdom of the crowd is the idea that an aggregated judgment of a group of individuals is often more accurate than the judgments of the individuals in the group (Davis-Stober et al., 2014; Galton, 1907; Surowiecki, 2004). The wisdom of the crowd is most often applied to continuous estimates or discrete choices, but has also been considered for ranking data (Lee et al., 2014).

Ranking or ordering is a common form of data in psychology, and provides a simple but informative way for people to express their knowledge. In the wisdom of the crowd setting, rankings have been used in the psychophysical task of aggregating people's perception of the weights of objects (Gordon, 1924). Rankings have been used to aggregate factual knowledge, such as people's ability to provide correct orders for domains like river lengths, city populations, and the ten commandments (Lee et al., 2011a, 2014), or to test people's ability to order items correctly in the NASA survival task (Hamada et al., 2020). Rankings have been used to aggregate people's episodic knowledge of event sequences (Steyvers et al., 2009). Finally, rankings have been used to aggregate people's predictions, such as the end-of-season order of sporting teams, the elimination sequence of contestants on a game show (Lee et al., 2011a, 2014), or the box-office success of movies (Selker et al., 2017).

A number of these wisdom of the crowds applications to ranking data have relied on the Thurstone cognitive model (e.g., Lee et al., 2011a, 2014; Selker et al., 2017; Steyvers et al., 2009). One common feature of all of these applications is that they involve complete rankings. Every individual is required to provide a ranking that includes every item under consideration. Thurstone models, however, are capable of considering partial or incomplete information. The original applications of the model to subjective perception of physical proximal stimuli (Thurstone, 1927a) and socially consensus opinions about offenses (Thurstone, 1927b) were both based on only pairwise judgments ranking two items. More recently, Böckenholt (1993, Section 9.1) developed general approaches for applying the Thurstone model to partial rankings, and Böckenholt (1992) presented an analysis to partial rankings of soft drink preferences.

The ability of the Thurstone model to deal with partial rankings seems especially useful in wisdom of the crowd applications. Consider three potential applications. The first involves the episodic memory eyewitnesses have for the sequence of events involved in a crime. Because of their vantage point, each eyewitness is able to provide a ranking of only the subset of events that they saw. The goal is to aggregate the partial rankings of the eyewitnesses to determine the overall true sequence of all of the events. The second potential application involves a set of sporting scouts evaluating potential players. Because each scout is limited to a geographical area or specializes in a specific type of player, they each evaluate only a subset of the players. The goal is to aggregate the partial rankings to form an overall ranking of the players on which to base recruitment decisions. The third potential application was introduced by Lee et al. (2015) and involves factual knowledge, such as identifying the top ten athletes in terms of personal wealth. It involves the situation in which people are not provided with a list of items, but must recall candidate items from memory. Because recall is be imperfect, different people's rankings will be based on different underlying subsets of the items, and thus their rankings need to be treated as being incomplete. The goal is to account for the failures of recall in aggregating the rankings.

These applications highlight two possible advantages of using partial rather than complete rankings. The first is to reduce the demand on peoples' time and effort. If there are a large number of items, it becomes difficult for people to provide complete rankings. The second advantage is that there is evidence that allowing people to contribute only the knowledge they are confident about can lead to better wisdom of the crowd performance. Kameda et al.'s (2022) review of information aggregation and collective intelligence argues that "opt-in and opt-out behavioral mechanisms can promote collective intelligence further in both consensus and combined decision making through capitalizing on individual heterogeneity in knowledge, skills and ability" (p. 354). As a specific example, in the context of true-or-false trivia questions, Bennett et al. (2018) show that crowd aggregates are more accurate when participants are allowed to "volunteer judgments" by selecting which questions they answer.

Given these potential applications and possible advantages, this article presents an approach to aggregating partial ranking to study the wisdom of the crowd. The approach relies on a Bayesian implementation of the Thurstone model, implemented using the JAGS graphical modeling language (Plummer, 2003). Previous work applying the Thurstone model to partial rankings (e.g., Böckenholt, 1992; Böckenholt, 1993) has not used Bayesian inference methods, and previous work using the Thurstone model with Bayesian methods (e.g., Johnson & Kuhn, 2013; Lee et al., 2014) has not involved applications to partial ranking data.

The structure of the remainder of this article is as follows. In the next section, we describe the basic assumptions of the Thurstone model of ranking, including its application to partial ranking data. We then consider applications to two different wisdom of the crowd problems: one involving the ranking the populations of ten US cities, and the other involving the chronological ordering of the first 44 US presidents. For both domains, we summarize previous result based on complete rankings, and present results based on new experiments using partial rankings. We consider two different methods for collecting partial rankings. One involves the experimenter selecting the subset of items each person ranks. The other allows the person to choose how many and which items they rank. We show that our Bayesian implementation of the Thurstone model applies naturally to aggregate both sorts of partial

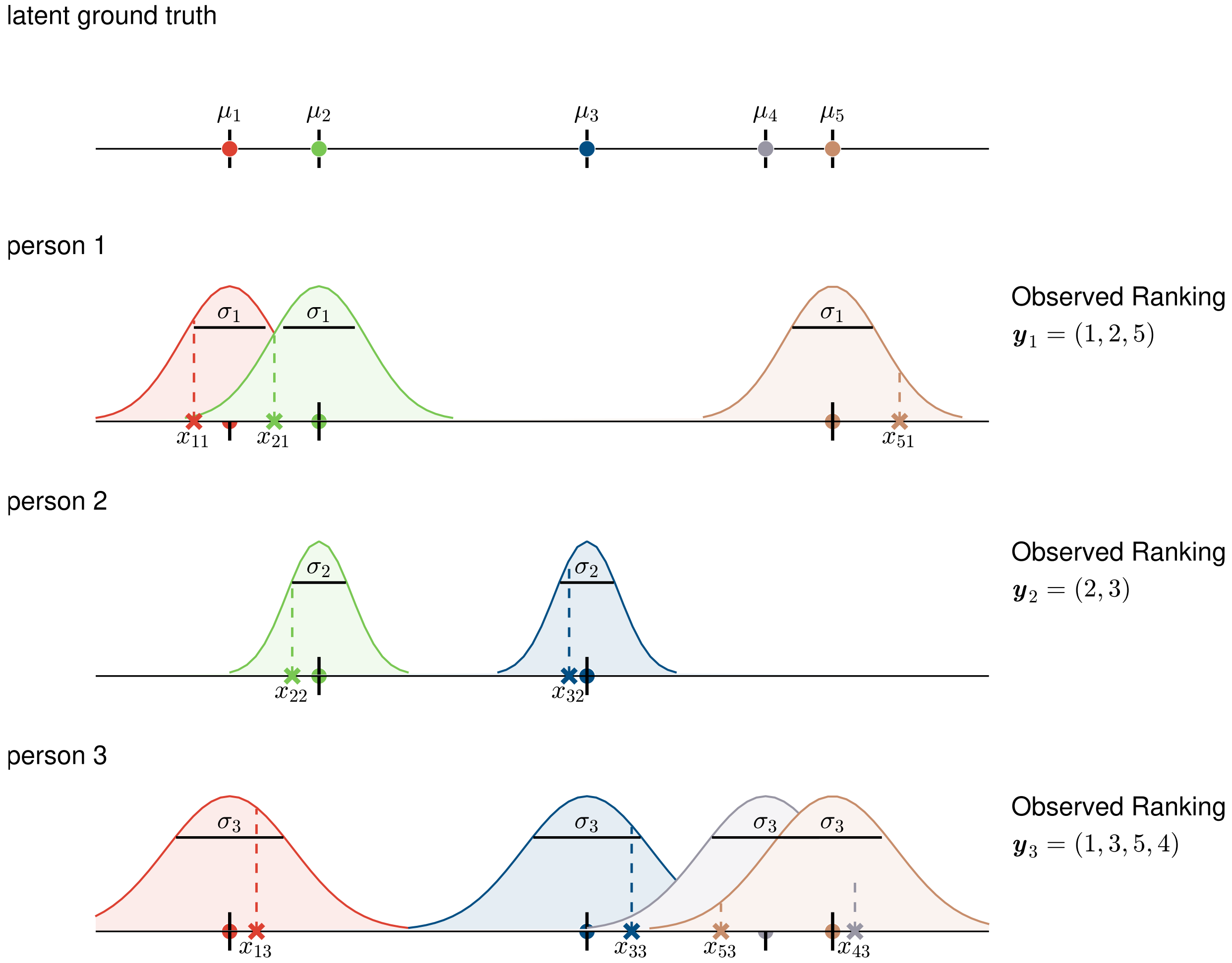

Figure 3.1: Conceptual representation of the Thurstone model applied to partial rankings.

data. We conclude with a discussion of how our approach raises, and can help answer, wisdom of the crowd research questions for situations in which people do not rank all of the items in a domain.

3.2 Thurstone Model for Partial Ranking

3.2.1 Model Assumptions

Figure 3.1 provides an overview of the assumptions of the Thurstone model as it applies to partial ranking data. The core assumption is that each item that can be ranked has a latent true location on an underlying unidimensional representation. In Figure 3.1, an example with five items is shown, and their latent grounds truths are represented by the parameters $\mu_1, \ldots, \mu_5$. The model assumes that each person knows the ground truth with some level of precision and, when asked to rank the item, draws mental samples representing their momentary understanding of each item with respect to the criterion. People's rankings are then determined by the ranking of the mental samples.

In Figure 3.1, Person 1 ranks only the first, second, and fifth items. Their mental samples, x_{11}, x_{21}, and x_{51} are drawn from Gaussian distributions centered on the ground truths μ_1, μ_2, and μ_5, with a standard deviation σ_1 that quantifies the precision of their knowledge. Formally, this means that

$$x_{ji} = \text{Gaussian}\big(\mu_j, \frac{1}{\sigma_i^2}\big), \tag{3.1}$$

where the Gaussian distribution is parameterized in terms of its mean and precision. The order of the mental samples for Person 1 is $x_{11} < x_{21} < x_{51}$, which leads to the observed partial ranking $\boldsymbol{y}_1 = (1, 2, 5)$. Person 2 ranks only the second and third items. Because $x_{22} < x_{32}$ their observed ranking is $\boldsymbol{y}_2 = (2, 3)$. Person 3 ranks all but the second item. Because $x_{13} < x_{33} < x_{53} < x_{43}$ their observed ranking is $\boldsymbol{y}_3 = (1, 3, 5, 4)$. The standard deviation σ_i can be interpreted as a measure of the expertise of the ith person: the smaller the standard deviation, the more likely it is the mental samples will be close to the ground truth. Figure 3.1 shows that $\sigma_2 < \sigma_1 < \sigma_3$, which means that Person 2 is relatively more expert than Person 1, but Person 3 is relatively less expert than both.

The Thurstone model does not provide an account of which items a person includes in their partial ranking. The strength of the model is its ability to make inferences about the ground truths $\boldsymbol{\mu}$ of the items and expertise $\boldsymbol{\sigma}$ of the individuals for any collection of partial rankings. The representational assumptions of the model, coupled with the simple decision processes assumed to generate rankings, provide a scaffolding that allows partial ranking data to be analyzed. Formally, the Thurstone model provides the likelihood $p\left(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\sigma}\right)$. Together with priors on $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, this allows Bayesian inferences to be made about the posterior $p\left(\boldsymbol{\mu}, \boldsymbol{\sigma} \mid \boldsymbol{y}\right)$. The posterior inferences for $\boldsymbol{\mu}$ correspond to the wisdom-of-the-crowd aggregate ranking, and the posterior inferences for $\boldsymbol{\sigma}$ correspond to measures of individual expertise.

The priors on $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ need to incorporate identifiability constraints. The same ranking data would be produced under translation and scaling of the underlying unidimensional representational. One way to impose constraints is allow initial uniform priors and then zero-center by requiring $\sum \mu_j = 0$ and unit-normalize by requiring $\langle \boldsymbol{\sigma} \rangle = 1$.

3.2.2 Model Implementation

Bayesian inference for Thurstone models generally requires computational methods based on Markov-chain Monte Carlo sampling. Custom samplers and libraries have been developed in computer science and statistics (e.g., Giles et al., 2018; Li et al., 2022), often with the ability to incorporate covariates relating to the expertise of individuals or the difficulties of items. Within psychology, Johnson & Kuhn (2013) pioneered a simpler implementation approach relying on the high-level graphical modeling language JAGS (Plummer, 2003).

A key assumption of the Thurstone model is that the order of the mental samples determines the rankings. This is an example of the statistical concept of censoring. JAGS implements censoring effectively through the use of its dinterval function. The general form of the dinterval function is $y_k = \text{dinterval}\big(x_k, \boldsymbol{b}\big)$, where $\boldsymbol{b} = (b_1, \ldots, b_m)$ is a vector of bounds that determine the mapping from the latent values in x to the censored observation. This means that

$$y_k = \begin{cases} 0 & \text{if } x_k \leq b_1 \\ 1 & \text{if } b_1 < x_k \leq b_2 \\ \ldots & \\ m-1 & \text{if } b_{m-1} < x_k \leq b_m \\ m & \text{if } b_m < x_k. \end{cases} \tag{3.2}$$

The dinterval function returns a lowest value of 0 for the first-ranked item. Ranking data, however, are usually represented as starting at rank 1. One way to accommodate this mismatch is to provide ranking data that start at 0 to the JAGS model.

The JAGS implementation developed by Johnson & Kuhn (2013), and used by others (e.g., Lee et al., 2014; Selker et al., 2017), applies dinterval and requires the separate calculation of the bounds for each position in a ranking, with placeholders below and above the first- and last-ranked items. A more direct use of dinterval was introduced by Lee & Ke (2022) in the context of modeling preferences in top-k lists. This approach uses the relevant mental samples themselves as the bounds, by sorting the mental samples:

$$y_k = \text{dinterval}\big(x_k, \text{sort}\,(\boldsymbol{x})\big). \tag{3.3}$$

In this formulation, the mental sample x_k is required to be in ranked position y_k among the mental samples. JAGS is able to make this joint censored inference, in which the mental sample for the current sample is constrained by bounds that correspond to other unknown mental samples.

To accommodate partial rankings, we extend this approach further by introducing an observed subset $\boldsymbol{s}$ that lists the items considered by the partial ranking. This leads to

$$y_k = \text{dinterval}\big(x_{s_k}, \text{sort}\,(\boldsymbol{x_s})\big). \tag{3.4}$$

The following JAGS script implements the Thurstone model for partial ranking data, in which `nPeople` consider a total of `nItems` and produce a total of `nRankings`. This allows for the same person to produce more than one ranking. The person who produced the ith ranking is `person[i]` and this partial ranking is given by `y[i,]`. The subset of items considered in this ranking is recorded in `set[i,]` and the number of items in the subset is `setN[i]`. The latent representation of the jth item is `mu[j]`, and the expertise of the ith person is `sigma[i]`. The observed inputs to the script are `nPeople`, `nItems`, `nRankings`, `person`, `set`, `setN`, and `y`. The parameters to be monitored are `mu` and `sigma`.

```
model{
 # Latent truth
 for (j in 1:nItems){
  muTmp[j] ~ dunif(-100,100)
  mu[j] = muTmp[j] - mean(muTmp)
 }
 # Expertise
 for (i in 1:nPeople){
  sigmaTmp[i] ~ dunif(0,1)
  sigma[i] = sigmaTmp[i]/mean(sigmaTmp)
 }
 # Data
 for (i in 1:nRankings){
  for (j in 1:nItems){
   x[i,j] ~ dnorm(mu[j],1/sigma[person[i]]^2)
  }
  for (j in 1:setN[i]){
   y[i,set[i,j]] ~ dinterval(x[i,set[i,j]],sort(x[i,set[i,1:setN[i]]]))
  }
 }
}
```

3.2.3 Concrete Example

As a concrete example of applying the model, consider the simple situation with three people and five items shown in Figure 3.1. The inputs to the model are

$$
\begin{aligned}
\texttt{nPeople} &= 3 \\
\texttt{nItems} &= 5 \\
\texttt{nRankings} &= 3 \\
\texttt{person} &= \begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \\
\texttt{y} &= \begin{bmatrix} 1 & 2 & - & - & 3 \\ - & 1 & 2 & - & - \\ 1 & - & 2 & 4 & 3 \end{bmatrix} \\
\texttt{subset} &= \begin{bmatrix} 1 & 2 & 5 & - \\ 2 & 3 & - & - \\ 1 & 3 & 4 & 5 \end{bmatrix} \\
\texttt{setN} &= \begin{bmatrix} 3 & 2 & 4 \end{bmatrix}.
\end{aligned}
$$

Note that the data in `y` are of the form that the entry in the jth column indicates the position in the partial ranking that the jth item was placed. The $-$ entries are needed to pad the `y` and `subset` matrices. These are implemented as missing, not available, or not-a-number values depending on the software being used to call JAGS.

All of the computational applications of the model reported in this article used eight chains. Most applications used 1000 burnin samples followed by 10,000 collected samples per chain, thinning by recording every 5th sample. The application to 194 partial rankings of the 44 US presidents required an additional 1000 samples burnin, and additional thinning by recording every 100th sample. With these settings, chains were observed to converge, based on visual inspection and the $\hat{R}$ statistic (Brooks & Gelman, 1998).

3.3 Application to US Cities Populations

Our first application of the model involves people's rankings of the population of ten US cities: New York, Los Angeles, Chicago, Houston, Phoenix, Philadelphia, San Antonia, San Diego, Dallas, and San Jose. We first summarize previously-published results involving complete rankings, then consider results for new experimental data involving both experimenter-selected and individual-selected approaches to partial rankings.

3.3.1 Complete Ranking

Lee et al. (2014) applied the Thurstone model to complete ranking data, in which every participant ranked every item. In one of their tasks, they had 142 participants provide a complete ranking of ten US city populations. Participants completed this task online using a drag and drop interface. The modeling results are shown in Lee et al. (2014, Figure 4). The key measure of wisdom the crowd performance is the difference between the model-inferred ranking and the true ordering. Lee et al. (2014) use Kendall's τ as a measure of performance. Kendall's τ counts the number of pairwise swaps needed to convert one ranking into another (Kendall, 1938; van Doorn et al., 2021). It provides an error measure, with $\tau = 0$ indicating a perfect estimate, and greater values of τ indicating successively less accurate estimates. Lee et al. (2014) report that the Thurstone model wisdom-of-the-crowd aggregate for the cities based on complete rankings has $\tau = 8$.

Lee et al. (2014) also report a predictive correlation between the Thurstone model's inferences about individual expertise and observed individual performance. Expertise is measured by

Table 3.1: The six different subsets of the ten US cities used to collect experimenter-selected partial rankings. Each participant was given just one subset, and provided a ranking of all six US cities in their subset.

Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6
New York	New York	Chicago	Los Angeles	New York	New York
Los Angeles	Los Angeles	Houston	Houston	Chicago	Houston
Phoenix	Chicago	Philadelphia	Phoenix	Houston	Phoenix
Philadelphia	San Antonio	San Antonio	San Diego	Phoenix	San Antonio
Dallas	Dallas	San Diego	Dallas	San Antonio	Dallas
San Jose	San Jose	San Jose	San Jose	San Diego	San Jose

the posterior of the σ_i parameter for each participant, and the accuracy of their observed rankings is again measured by Kendall's τ. Since σ_i is a standard deviation on the precision of mental samples, smaller values correspond to greater expertise. Similarly, smaller Kendall's τ values correspond to greater accuracy. Lee et al. (2014) report a correlation of $r = 0.79$ between the individuals' inferred expertise and their accuracy, meaning that participants inferred to be more expert tend to provide more accurate rankings. Lee et al. (2014) emphasize that this result genuinely involves *predictions* about expertise and performance, since the model never has access to the ground truth, and hence has no information about participant performance.

3.3.2 Experimenter-Selected Partial Ranking

Participants

A total of 62 participants were recruited on Prolific (`www.prolific.com`). All of the Prolific participants were US citizens. Their median age was 32 years.

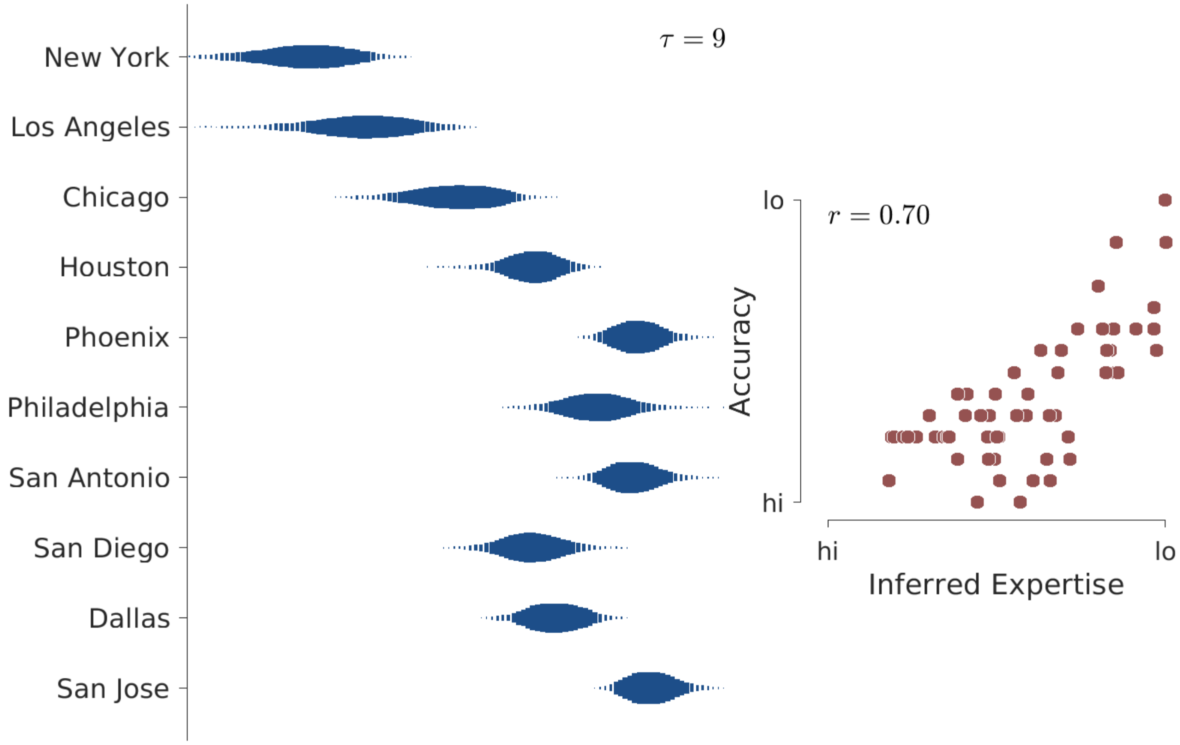

Figure 3.2: The main panel shows the model's inferred marginal posterior distributions for the μ_j of each city, based on experimenter-selected partial rankings. The inset panel shows the relationship across participants between their inferred expertise, as measured by the posterior mean of σ_i, and their observed accuracy, as measured by Kendall's τ.

Procedure

Participants provided rankings for one of six different subsets, each of which contained six cities. The six subsets were constructed beforehand and are listed in Table 3.1. They were chosen at random, subject to two constraints. Each city needed to occur in at least two subsets, and every pair of cities needed to occur together in at least one subset. The task was completed online using the "Rank Order" question type provided by Qualtrics (`www.qualtrics.com`). Due to an experimenter error, one subset was completed by twelve participants while the other five had ten participants.

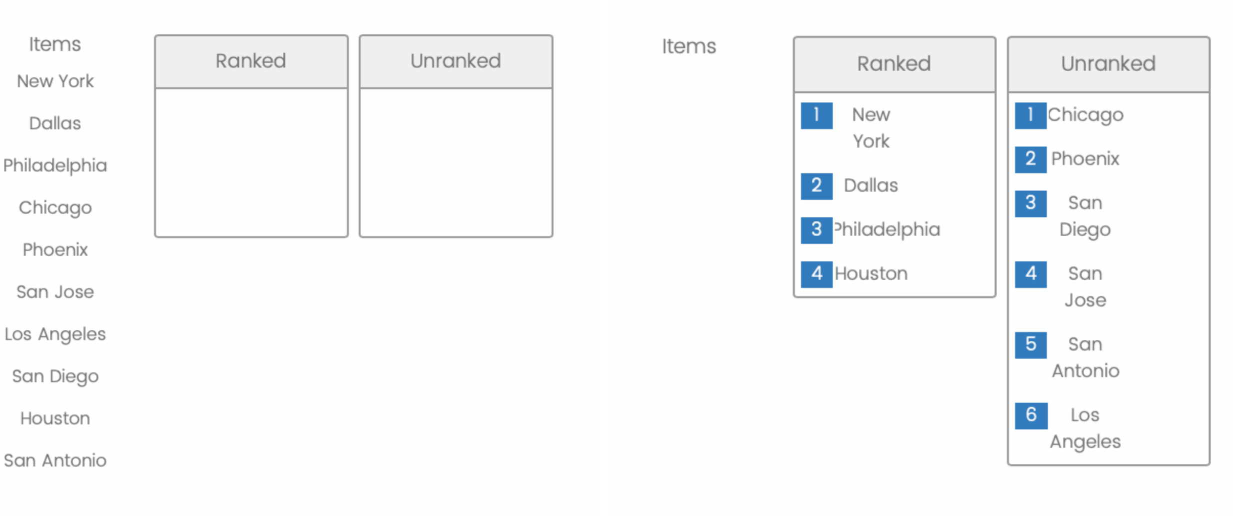

Figure 3.3: Screenshots of the Qualtrics interface for the US city population task in the individual-selected partial ranking experiment. The left panel shows the initial state of the interface. The right panel shows an illustrative participant response.

Results

The results of applying the Thurstone model to the partial ranking data are show in Figure 3.2. The ten cities are ordered top to bottom in terms of their population from largest to smallest. The violin plots show the marginal posterior distributions of the ground truth μ_j for each city. The model's inferred ranking is determined by the means of these posterior distributions: New York, Los Angeles, Chicago, San Diego, Houston, Dallas, Philadelphia, San Antonio, Phoenix, and San Jose. This wisdom-of-the-crowd aggregate ranking is $\tau = 9$ pairwise swaps from the true ordering. The inset scatter plot in Figure 3.2 compares inferred individual expertise to performance, and shows a correlation of $r = 0.70$.

3.3.3 Individual-Selected Partial Ranking

Participants

A total of 194 undergraduate students in a psychology class at the University of California Irvine participated for class credit.

Procedure

Participants provided partial rankings in a setting in which they could choose *how many* and *which* cities they ranked. The left panel of Figure 3.3 provides an example of the Qualtrics interface, developed using the "Pick, Group, & Rank" question type. Participants were required to drag and drop all ten items into either the ranked or unranked box. They were asked to rank a minimum of two items, but otherwise only rank the items that they were confident they could rank correctly. The right panel of Figure 3.3 provides an illustrative participant response in which New York, Dallas, Philadelphia and Houston are ranked, in that order, but the other six cities are not ranked.

Results

Figure 3.4 orders participants from left to right in terms of how many cities they ranked, and the cities in terms of the true order of their population from top to bottom. The blue lines in the main axes denote the cities that a participant ranked. The right margin bar graph shows how often each city was ranked. Some cities, like New York and Los Angeles, were ranked more often than others. The top margin bar graph shows how many cities each participant ranked, beginning with ten and ending with two. Green bars indicate correct rankings with $\tau = 0$, while yellow bars indicate incorrect rankings with $\tau > 0$. The top

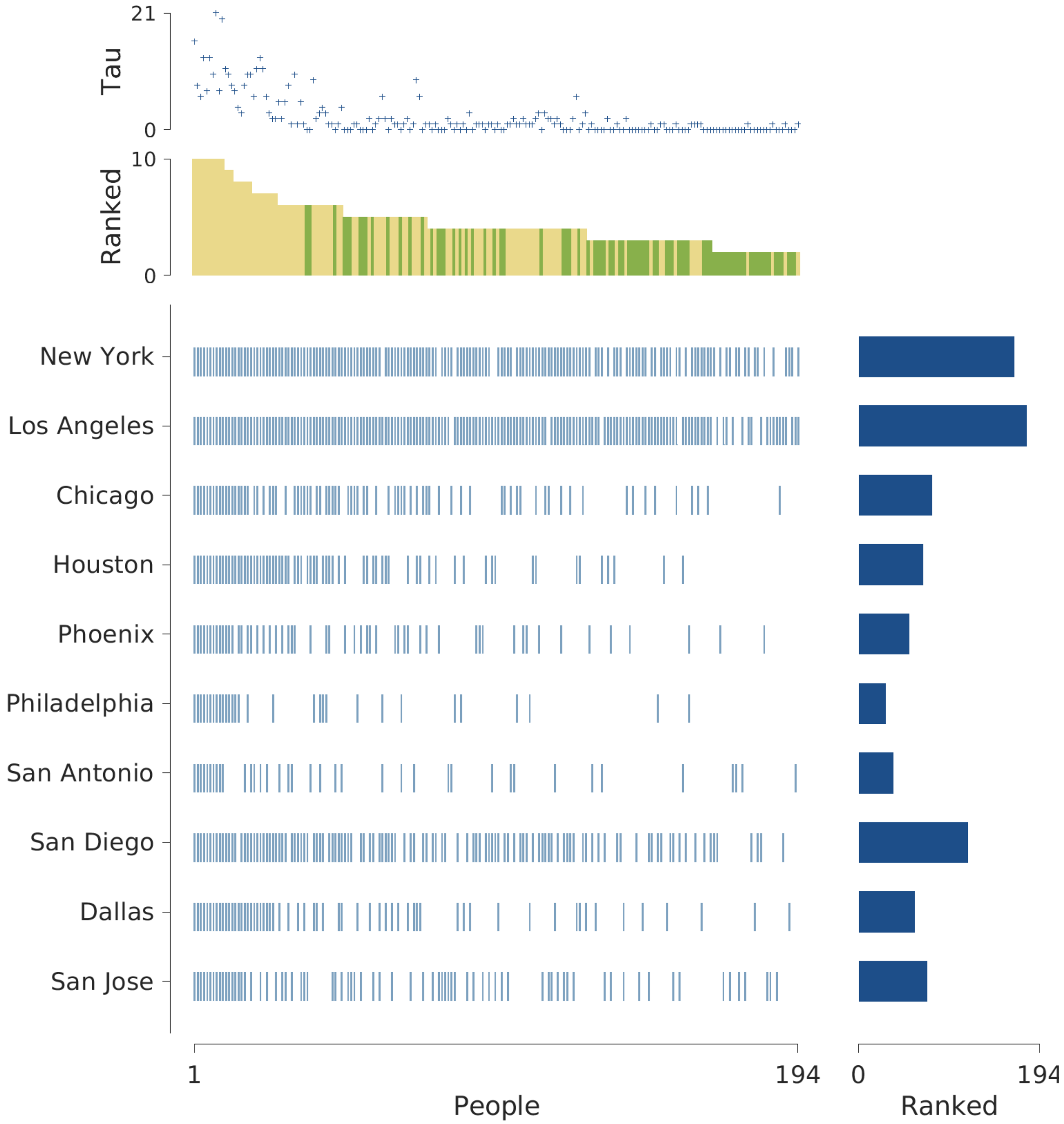

Figure 3.4: The main axes shows participants ordered from left to right according to how many cities they ranked, and the cities' true population ordering from top to bottom. The blue lines in the main axes denote the cities a participant ranked. The right margin bar graph shows how often every city was ranked. The top margin bar graph shows how many cities a participant ranked. Green and yellow bars indicate correct and incorrect rankings. The top margin with crosses shows the τ measure for each participant.

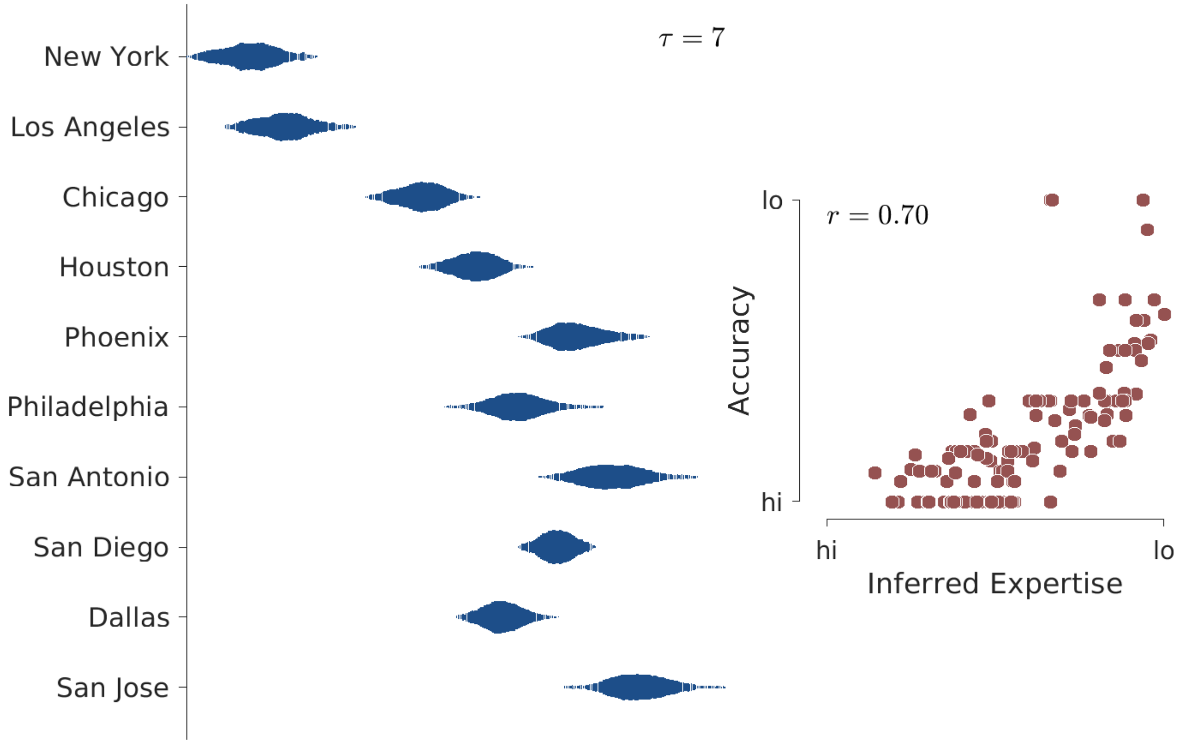

Figure 3.5: The main panel shows the model's inferred marginal posterior distributions for the μ_j of each city, based on individual-selected partial rankings. The inset panel shows the relationship across participants between their inferred expertise, as measured by the posterior mean of σ_i, and their observed accuracy, as measured by Kendall's τ.

margin with crosses shows the τ error measure for each participant.

The Thurstone model results for the individual-selected partial rankings are shown in Figure 3.5. The wisdom-of-the-crowd aggregate ranking is $\tau = 7$ pairwise swaps from the true ordering, and a correlation of $r = 0.70$ between the individuals' inferred expertise and their accuracy.

The results of the Thurstone model with both the experimenter-selected and individual-selected partial rankings are very similar to the results based on complete rankings. It does not make sense to compare the three sets of results closely, because they do not come from a within-participants design, and the numbers of participant and recruitment populations are very different. The results do suggest, however, that partial ranking data can allow the Thurstone model to make similarly accurate inferences as complete ranking data about the wisdom-of-the-crowd aggregate and individual expertise.

3.4 Application to US Presidents Chronology

Our second application of the model involves people's rankings of the chronological order of the first 44 US presidents. This provides a test of the ability of our approach to scale to larger domains. Once again, we first summarize previously-published results involving complete rankings, before considering new experimental data using experimenter-selected and individual-selected partial rankings.

3.4.1 Complete Ranking

Lee et al. (2014) asked 26 participants to rank the first 44 US presidents in chronological order. These participants ranked the presidents in person using physical cards. Lee et al. (2014, Figure 3) report that the Thurstone model wisdom-of-the-crowd aggregate ranking is $\tau = 37$ pairwise swaps from the true ordering, and a correlation of $r = 0.95$ between inferred individuals' expertise and their observed performance.

3.4.2 Experimenter-Selected Partial Ranking

Participants

The same 62 Prolific participants who did US cities experimenter-selected partial ranking task completed the US presidents experimenter-selected partial ranking task.

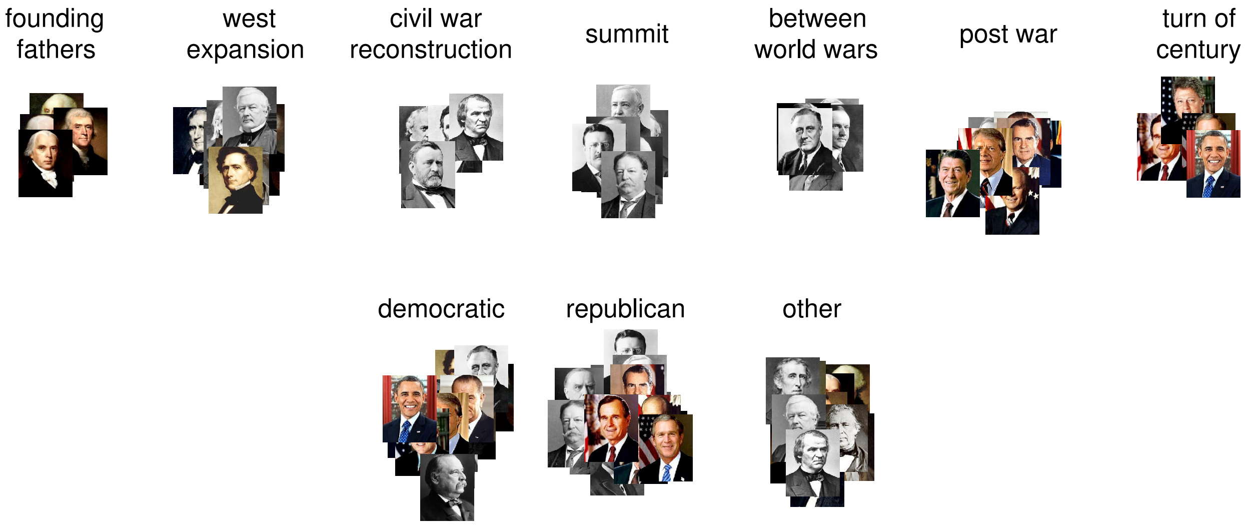

Figure 3.6: Subsets of 44 US presidents based on time period (top row) or political party affiliation (bottom row). Time period divided the presidents into seven subsets, while political affiliation divided them into three subsets.

Procedure

There are various ways in which experimenter-selected partial ranking tasks can be constructed. The US cities task generated random overlapping subsets with the same number of items. For some ranking tasks, however, there may be more meaningful ways to construct subsets. The US presidents can be naturally grouped in terms of different time periods ("founding fathers," "civil war reconstruction," and so on) or in terms of party affiliation ("Democratic," "Republican," or "other"). These natural groupings may target the diversity of niche expertise that different people have, and improve the knowledge provided by individuals. We appealed to this motivation for the experimenter-selected partial rankings of the presidents.

Participants completed either the time period or party affiliation condition. Figure 3.6 shows the subsets for each of the two conditions. The time period condition grouped the presidents into seven subsets, while the party affiliation condition grouped them into three subsets. Participants in both conditions considered all 44 presidents, but they did so by ranking the presidents within each of the subsets for their condition, completing one subset at a time.

This approach is one in which the same person provides multiple partial rankings. The data were again collected using the "Rank Order" question type provided by Qualtrics.

Results

The Thurstone model results for the experimenter-selected partial ranking tasks are shown in Figure 3.7. The wisdom-of-the-crowd aggregate ranking is $\tau = 55$ pairwise swaps from the true ordering, and a correlation of $r = 0.79$ between the individuals' inferred expertise and their accuracy.

3.4.3 Individual-Selected Partial Ranking

Participants

The same 194 undergraduate students who did the US cities individual-selected partial ranking task completed the US presidents individual-selected partial ranking task.

Procedure

Participants provided partial rankings for this task in an identical fashion to the US cities individual-selected partial rankings task. Using the "Pick, Group, & Rank" question type on Qualtrics, they dragged all 44 presidents into either the ranked or unranked box. They were required to rank a minimum of two items, and beyond that add only the items they were confident they could correctly include in their ranking.

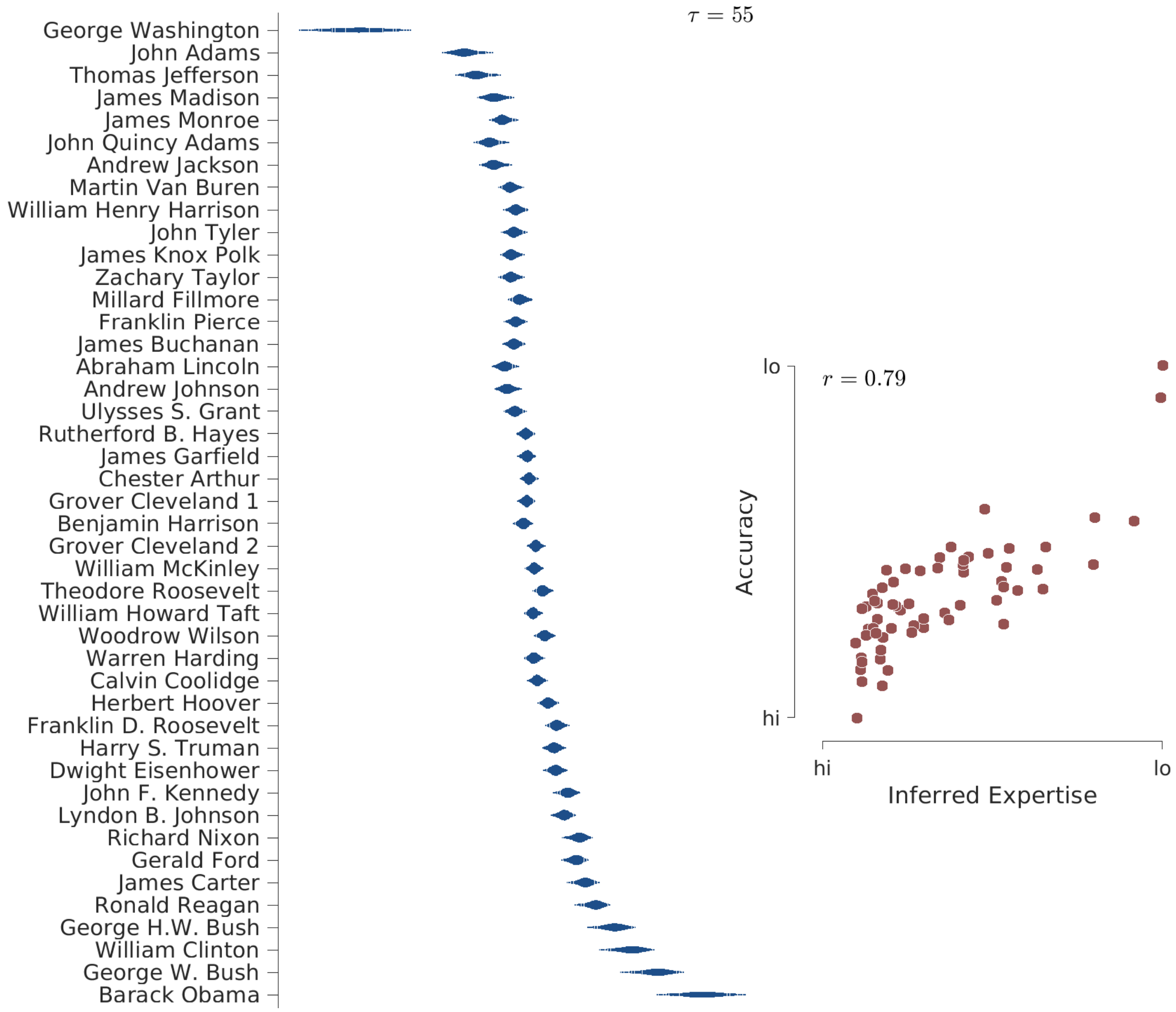

Figure 3.7: The main panel shows the model's inferred marginal posterior distributions for the μ_j of each city, based on experimenter-selected partial rankings. The inset panel shows the relationship across participants between their inferred expertise, as measured by the posterior mean of σ_i, and their observed accuracy, as measured by Kendall's τ.

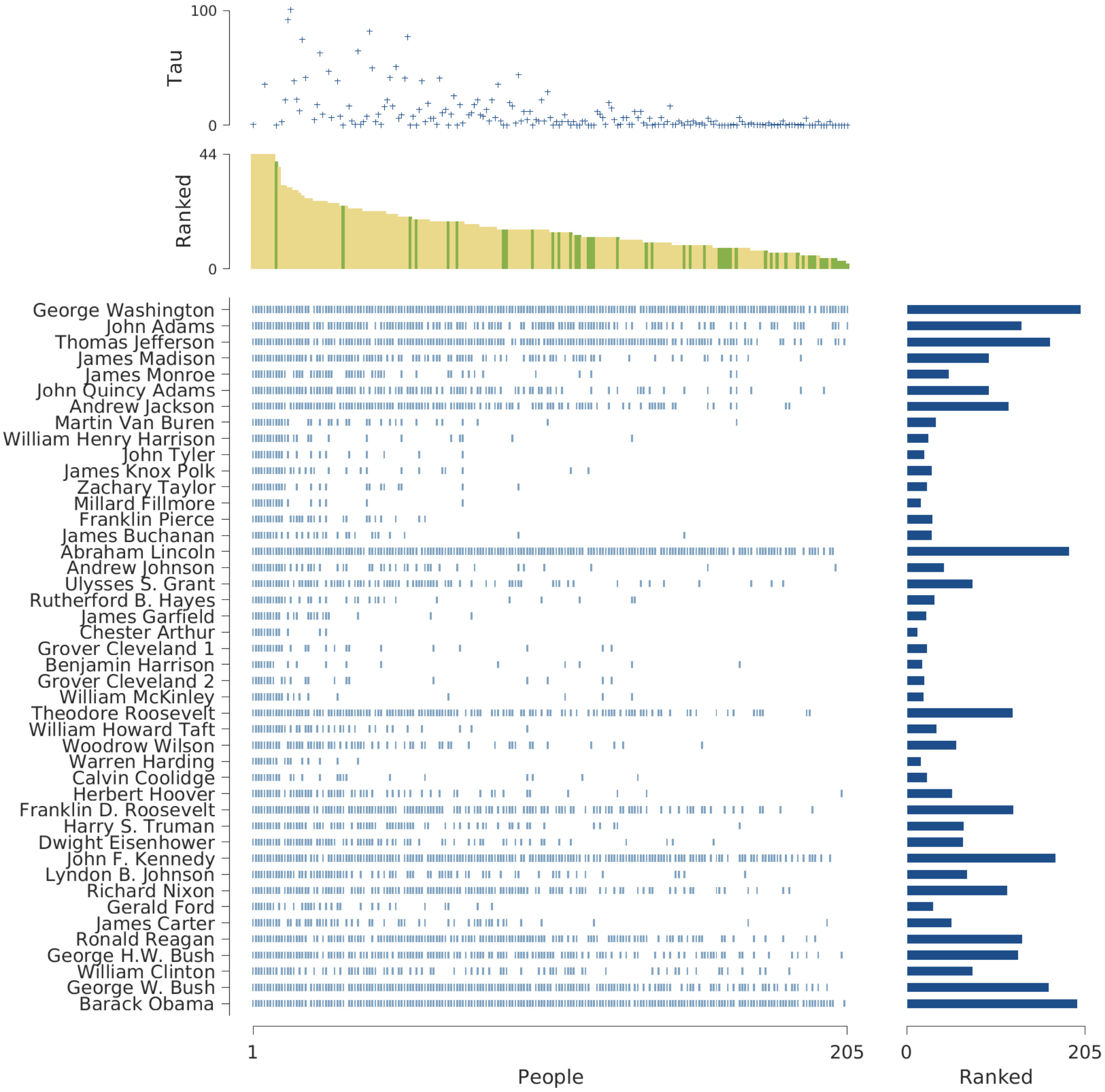

Figure 3.8: The main axes shows participants ordered from left to right according to how many presidents they ranked, and the presidents' true ordering from top to bottom. The blue lines in the main axes denote the presidents a participant ranked. The right margin bar graph shows how often every president was ranked. The top margin bar graph shows how many presidents a participant ranked. Green and yellow bars indicate correct and incorrect rankings. The top margin with crosses shows the τ measure for each participant.

Results

Figure 3.8 provides greater detail about how individual participants ranked the presidents. Some presidents like George Washington, Abraham Lincoln, and Barack Obama were included in almost all participants' rankings, while other presidents like Chester Arthur and Millard Fillmore appeared infrequently. Only eight participants chose to rank all of the presidents. The median number of presidents ranked is 14, and 90% of participants ranked fewer than 26 presidents.

The Thurstone model results for the individual-selected partial rankings are presented in Figure 3.9. The wisdom-of-the-crowd aggregate ranking is $\tau = 2$ pairwise swaps from the true ordering, and a correlation of $r = 0.85$ between the individuals' inferred expertise and their accuracy.

In this application, the results of the Thurstone model with individual-selected partial rankings are impressively accurate, and significant better than the experimenter-selected and complete rankings. Of course, there are still important differences in the participant pools, and the larger number of participants for individual-selected partial rankings may explain some or all of the improvement. The results do, however, provide an encouraging example of the potential of individual-selected partial rankings.

3.5 Discussion

The two applications demonstrate that our JAGS implementation of the Thurstone model can be applied to partial ranking data. This capability allows a number of problems related to the wisdom of the crowd to be studied. One central question is the relative merits of collecting complete rankings versus some form of partial ranking. Our applications considered both experimenter-selected and individual-selected partial rankings. In the experimenter-selected

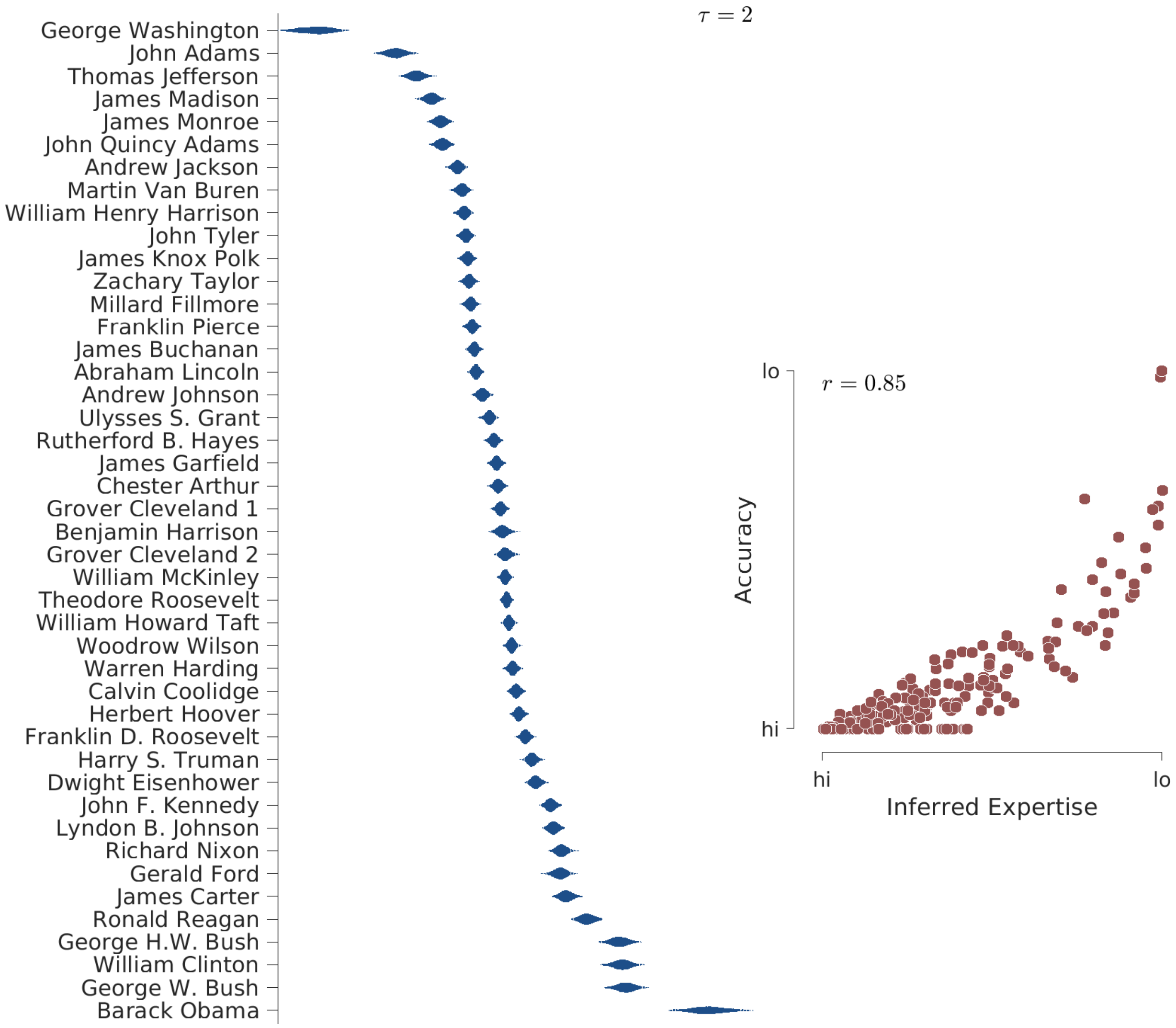

Figure 3.9: The main panel shows the model's inferred marginal posterior distributions for the μ_j of each president, based on individual-selected partial rankings. The inset panel shows the relationship across participants between their inferred expertise, as measured by the posterior mean of σ_i, and their observed accuracy, as measured by Kendall's τ.

situation, how the subsets should be chosen is an interesting research question. In the individual-selected situation, what instructions people should be given in terms of balancing their completeness and accuracy is also an interesting research question.

A more ambitious extension involves allowing the same person to choose and rank multiple, possibly overlapping, subsets. For example, a person may be uncertain whether New York or Los Angeles has the greater population, and be uncertain whether Houston or Chicago has the greater population, but nonetheless be certain both of the first pair have greater populations than both of the second pair, and also be that certain all four cities have greater populations than Philadelphia. This knowledge could be expressed in four partial rankings: New York > Houston > Philadelphia, New York > Chicago > Philadelphia, Los Angeles > Houston > Philadelphia, and Los Angeles > Chicago > Philadelphia. To accommodate this extension, the JAGS implementation has to allow for there to be more partial rankings than people, and to apply the person-specific expertise σ_i to every ranking from the same person. These are straightforward extensions and a script implementing the multiple partial rankings situation is provided in the supplementary information.

Another research question is whether the use of the Thurstone model-based method improves wisdom-of-the-crowd aggregates over statistical methods. For complete rankings, there is evidence that the Thurstone approach improves on Borda aggregation (e.g., Lee et al., 2011a, 2014). This occurs because the model's inferences about relative expertise allow it to give greater weight to experts, in distinction to the Borda count which weights all individuals equally. Various modified Borda count and other statistical aggregation methods have been develop for partial rankings (Chen et al., 2022; Goddard, 1983; Herrero & Villar, 2021; Ju et al., 2015). It would be interesting to compare the performance of model-based and statistical aggregates systematically and thoroughly. As we mentioned, the results of the individual-selected partial rankings for the US presidents are very encouraging.

Several of our motivating examples for partial ranking aggregation involved real-world problems like eyewitness testimony and sport scouting. A feature of our approach is that it can, in principle, accommodate any number of people, items, and any pattern with which those people partially rank the items. This is important for naturally occurring data associated with real-world problems that lack experimental control. The use of Bayesian methods are particularly important in this regard. There are many possibilities in which there has to be significant uncertainty about the aggregate ranking, and coherent Bayesian representation of that uncertainty is required. As a simple example, consider a situation in which one set of scouts ranks players from one team, and a different set of scouts ranks players from another team. These partial rankings will provide information for within-team aggregation, but there is no information for between-team comparison of players. The result from our method would be a joint posterior over the $\boldsymbol{\mu}$ parameters that captured this uncertainty, providing posterior mass to every possible between-team combination consistent with the within-team rankings.

The Bayesian framework also allows the introduction of informative priors (Lee & Vanpaemel, 2018). One possibility in wisdom-of-the-crowd ranking applications is to use priors based on logical constraints. For example, in eyewitness event reconstruction, order-restricted priors could be placed on $\boldsymbol{\mu}$ to force logically connected events to follow the required sequence. Altmann (2003) considers examples like this from people's memory of events on September 11. For example, a World Trade Center tower falling can only happen after it was hit by a plane. Order constraints on the appropriate $\boldsymbol{\mu}$ parameters formalize this logic, and allow the model to infer the remainder of the sequence from people's ranking judgments.

Another extension would be to incorporate covariate information about people or items. This approach has been pioneered by Johnson & Kuhn (2013). It is likely to be especially useful in the context of real-world applications. It would allow, for example, the inferred expertise of scouts to be regressed on their years of experience, or the expertise of eyewitnesses to be regressed on assessments of the acuity of their vision or the reliability of their memory.

Finally, Thurstone partial rankings could be applied beyond wisdom-of-the-crowd settings. Lee & Ke (2022) use top-k variants to explore the structure of people's preferences. These no longer have a ground truth, but inferences about the underlying ranking corresponds to something like a cultural consensus (Anders & Batchelder, 2015; Romney et al., 1987). As a concrete example, consider the lists of "Good Reads" provided by book lovers. These are partial rankings of all possible books. They are not the top-k lists considered by Lee & Ke (2022), because a top-k list would require every possible book to be considered. Instead, they are partial rankings based on the subset of books an individual person has read (or otherwise decides to consider in constructing their ranking). This is exactly the type of partial ranking structure that our model can aggregate.

Ranking data are a ubiquitous way by which people express knowledge and an important case for studying the wisdom of the crowd. In many situations, it is impractical or undesirable for people to rank all of the items being considered. In this article, we have developed, implemented, and demonstrated a Bayesian method for using the Thurstone to aggregate partial rankings. We hope it serves as a useful tool for studying and applying the wisdom of the crowd.

3.6 Publication Note

This chapter has been submitted has a manuscript for publication.

CONCLUSION

The central focus of this dissertation is on preserving and better capturing the diversity present in the crowd. This is done by using cognitive modeling that considers contextualized expertise and by using different task designs to gather more informed or multiple estimates from the individuals within the crowd. The two applications that I use to illustrate these points are a task that requires spatial knowledge, Chapters 1 and 2, and a subset ranking task, Chapter 3. This work complements the existing discussion in the literature about how to construct a *better* crowd. All three chapters discuss ways of collecting multiple estimates from the same individual, and, with the exception of Chapter 2, how the wisdom of the crowd estimates can be further compared to wisdom of the crowd within estimates. I see this work as proof of concept for how to use the wisdom of the crowd within in conjunction with cognitive modeling to produce improved wisdom of the crowd estimates.

Simply asking a participant for a single estimates or just getting one value from them does not necessarily allow them to express their knowledge fully. Chapter 1 explores how asking participants for the same information in different frames may allow for them to break the information they know into component parts. I found that there was a framing effect: participants responded more conservatively in the absent frame than they did in the present frame. By constructing the task to take advantage of this framing effect, a participant can supply multiple estimates, and these estimates can then be used to generate wisdom of the crowd within estimates. Chapter 2 has participants supply two different types of estimates that get at different ways of expressing where they think a particular US city is located. Participants were asked to give a point estimate and a region that the city should be contained in. These two measures gave different ways to evaluate their performance and accuracy, while also evaluating whether the wisdom of the crowd estimates improved by using one or both of these measures. I further explored whether collecting additional demographic information about the states they were familiar with was helpful in better contextualizing their expertise. While this was not the case, it may be that asking participants what states they were familiar with in a more explicit fashion (i.e., "What states have you lived in within the last five years?") would be more helpful in shining light on why the individual-by-city expertise was a helpful parameter to include in the cognitive model. Chapter 3 provides a framework for how eliciting different rankings from the same individual can be accounted for using a cognitive model. The complete and experimenter-selected rankings were intentionally constructed so as to get rankings that exhaustively compared the items, and a concern of using the individual-selected partial rankings was that the model's performance would be impacted by having incomplete pairwise comparisons. However, this was not the case; the model's ranking for the 44 US presidents only required 2 pairwise swaps to get to the true ordering. Thus, it appears that the Thurstone cognitive model is robust in dealing with the missing pairwise comparisons.

Whether in the context of spatial knowledge or ranking tasks, eliciting multiple estimates from the same individual makes questions about expertise all the more interesting. An unsurprising and general result from Chapter 1 was that participants found the US states condition was easier than the African countries condition. Extending this result from the condition-level to that of the item-level, the work in Chapter 2 more specifically addresses how item difficulty might specifically impact spatial knowledge. The cognitive model described in Chapter 2 has a multivariate distribution with a covariance matrix. Its covariance matrix in Equation 2.4 has three components in its variance. These terms correspond to the city's difficulty, an individual's overall expertise, and an individual-by-city expertise. The results from comparing the various cognitive model-based wisdom of the crowd estimates demonstrate that there is some trade-off between including or excluding these terms. Generally though, the model-based estimates improve when individual expertise is considered. There is not a clear result about how city difficulty and individual-by-city expertise interact from the two datasets analyzed. Our hypothesis was that the individual-by-city difficulty and the city difficulty might have some relation with the demographic information about how familiar individuals were with particular states. There was not much difference between the posterior means of the individual-by-city expertise, β_{ij}, when separated by which cities were within familiar states and which were not. In all, we were unable to determine exactly what is responsible for why including these parameters improves the model-based wisdom of the crowd estimates with the data that we have from participants. We think that a useful analysis to answer this question might instead consider how these parameters interact with geographic constraints or population density. Cities closer to the coasts appear to be less difficult, which could be due to how those cities and states have obvious boundaries on one side or that participants in general had more knowledge about locations with larger populations.

Another consideration to make is how the multiple estimates from the same individual should be combined. Wisdom of the crowd within performs the best when the estimates from the individuals are as independent as possible (Herzog & Hertwig, 2009; Vul & Pashler, 2008). One way to do this is to manipulate the way the question is being asked, as was done in Chapter 1. Instead of asking the same question at different times or having participants combine estimates themselves, we focused on having participants answer a question in different ways. From the research on the framing effect, it is known that participants do not treat responses to inclusion and exclusion frames equivalently. Participants tend to be risk seeking in their responses when asked in an inclusion framework, and risk averse in the exclusion framework. Thus, for the tile map selection task in Chapter 1, we were able to elicit a framing effect by using the present and absent framings, comparable to the inclusion and exclusion framings, which can be seen in the behavioral analyses for this chapter. The experimenter-selected partial ranking task in Chapter 3 similarly had participants rank US presidents by time period or political party affiliation. For ranking tasks like these, however, we thought that participants being able to select the items they ranked would most improve their responses. Bennett et al. (2018) and Kameda et al. (2022) found and argue that allowing participants to opt-in or "volunteer" responses leads to greater accuracy in their responses by better facilitating nuanced individual responses.

Taken together, these results provide insight into how to increase or preserve diversity in the wisdom of the crowd. Diversity was introduced as the information that an individual uniquely knows, and to capture this information requires considering how to elicit varied estimates from individuals that better encapsulates what they think and believe. Furthermore, diversity also refers to how to best leverage individual expertise and upweight those who know more. Cognitive modeling is invaluable in accomplishing this as the models discussed in Chapters 2 and 3 have latent parameters related to individual expertise embedded within them.

The obvious future steps for this work would be developing a cognitive model for Chapter 1 and comparing the model-based wisdom of the crowd and model-based wisdom of the crowd within estimates. The cognitive model from this task will probably be conceptually similar to our cognitive model in Chapter 2 (see Montgomery et al., 2024) and Mayer & Heck (2023)'s two-dimensional cultural consensus theory cognitive model. Another interesting expansion of Chapter 3 would be to have participants supply multiple participant-selected partial rankings of the same data set.

Bibliography

Abi-Zeid, I. & Frost, J. R. (2005). SARPlan: A decision support system for Canadian Search and Rescue operations. *European Journal of Operational Research*, *162*(3), 630–653, https://doi.org/10.1016/j.ejor.2003.10.029.

Altmann, E. M. (2003). Reconstructing the serial order of events: A case study of September 11, 2001. *Applied Cognitive Psychology*, *17*(9), 1067–1080, https://doi.org/10.1002/acp.986.

Anders, R. & Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, *56*(6), 452–469, https://doi.org/10.1016/j.jmp.2013.01.004.

Anders, R. & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*(1), 151–181, https://doi.org/10.1007/s11336-013-9382-9.

Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, *61*, 1–13, https://doi.org/10.1016/j.jmp.2014.06.001.

Armstrong, J. S. (2001). Combining Forecasts. In A. J. S (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Springer.

Atanasov, P., et al. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, *63*(3), 691–706, https://doi.org/10.1287/mnsc.2015.2374.

Aydin, B. I., Yilmaz, Y. S., Li, Y., Li, Q., Gao, J., & Demirbas, M. (2014). Crowdsourcing for multiple-choice question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, *28*(2), 2946—2953, https://doi.org/10.1609/aaai.v28i2.19016.

Bennett, S. T., Benjamin, A. S., Mistry, P. K., & Steyvers, M. (2018). Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. *Computational Brain & Behavior*, *1*, 90–99, https://doi.org/10.1007/s42113-018-0006-4.

Böckenholt, U. (1992). Thurstonian representation for partial ranking data. *British Journal of Mathematical and Statistical Psychology*, *45*(1), 31–49, https://doi.org/10.1111/j.2044-8317.1992.tb00976.x.

Böckenholt, U. (1993). Applications of Thurstonian models to ranking data. In *Probability models and statistical analyses for ranking data* (pp. 157–172). Springer.

Boon, M. (2012). Predicting elections: A 'wisdom of crowds' approach. *International Journal of Market Research*, *54*(4), 465–483, https://doi.org/10.2501/IJMR-54-4-465-483.

Breivik, Ø., Allen, A. A., Maisondieu, C., & Olagnon, M. (2013). Advances in search and rescue at sea. *Ocean Dynamics*, *63*, 83–88, https://doi.org/10.1007/s10236-012-0581-1.

Brooks, S. P. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455, https://doi.org/10.1080/10618600.1998.10474787.

Bruce, R. S. (1935). Group judgments in the fields of lifted weights and visual discrimination. *The Journal of Psychology*, *1*(1), 117–121, https://doi.org/10.1080/00223980.1935.9917245.

Budescu, D. V. & Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280, https://doi.org/10.1287/mnsc.2014.1909.

Butler, D., Butler, R., & Eakins, J. (2021). Expert performance and crowd wisdom: Evidence from English Premier League predictions. *European Journal of Operational Research*, *288*(1), 170–182, https://doi.org/10.1016/j.ejor.2020.05.034.

Chen, W., Zhou, R., Tian, C., & Shen, C. (2022). On top-k selection from m-wise partial rankings via Borda counting. *IEEE Transactions on Signal Processing*, *70*, 2031–2045, https://doi.org/10.1109/TSP.2022.3167159.

Christiansen, J. D. (2007). Prediction markets: Practical experiments in small markets and behaviours observed. *The Journal of Prediction Markets*, *1*(1), 17–41, https://doi.org/10.5750/jpm.v1i1.418.

Council, N. R. (2013). *Future U.S. Workforce for Geospatial Intelligence*. Washington, DC: National Academies Press.

Da, Z. & Huang, X. (2019). Harnessing the wisdom of crowds. *Management Science*, *66*(5), 1847–1867, https://doi.org/10.1287/mnsc.2019.3294.

Danileiko, I. & Lee, M. D. (2018). A model-based approach to the wisdom of the crowd in category learning. *Cognitive Science*, *42*(S3), 861–883, https://doi.org/10.1111/cogs.12561.

Davis-Stober, C. P., Budescu, D. V., Broomell, S. B., & Dana, J. (2015). The composition of optimally wise crowds. *Decision Analysis*, *12*(3), 130–143, https://doi.org/10.1287/deca.2015.0315.

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79–101, https://doi.org/10.1037/dec0000004.

Deary, I. J. (2020). *Intelligence: A very short introduction.* Oxford University Press, 2nd edition.

Drew, T., Vo, M. L., Olwal, A., Jacobson, F., Seltzer, S. E., & Wolfe, J. M. (2013). Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision*, *13*(10), 1–13, https://doi.org/10.1167/13.10.3.

Farnsworth, P. R. & Williams, M. F. (1936). The accuracy of the median and mean of a group of judgments. *The Journal of Social Psychology*, *7*(2), 237–239, https://doi.org/10.1080/00224545.1936.9921664.

Fiechter, J. L. & Kornell, N. (2021). How the wisdom of crowds, and of the crowd within, are affected by expertise. *Cognitive Research: Principles and Implications*, *6*(5), https://doi.org/10.1186/s41235-021-00273-6.

Friedman, A., Brown, N. R., & Mcgaffey, A. P. (2002a). A basis for bias in geographical judgments. *Psychonomic Bulletin & Review*, *9*(1), 151–159, https://doi.org/10.3758/bf03196272.

Friedman, A., Kerkman, D. D., & Brown, N. R. (2002b). Spatial location judgments: A cross-national comparison of estimation bias in subjective North American geography. *Psychonomic Bulletin & Review*, *9*(3), 615–623, https://doi.org/10.3758/bf03196321.

Friedman, A., Kerkman, D. D., Brown, N. R., Stea, D., & Cappello, H. M. (2005). Cross-cultural similarities and differences in North Americans' geographic location judgments. *Psychonomic Bulletin & Review*, *12*(6), 1054–1060, https://doi.org/10.3758/bf03206443.

Friedman, A., Mohr, C., & Brugger, P. (2012). Representational pseudoneglect and reference points both influence geographic location estimates. *Psychonomic Bulletin & Review*, *19*, 277–284, https://doi.org/10.3758/s13423-011-0202-x.

Fu, L., Lee, L., & Danescu-Niculescu-Mizil, C. (2017). When confidence and competence collide: Effects on online decision-making discussions. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1381–1390).

Fu, L., Wang, A. Z., & Danescu-Niculescu-Mizil, C. (2020). Confidence Boost in Dyadic Online Teamwork: An Individual-Focused Perspective. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14 (pp. 197–208).

Galton, F. (1907). Vox populi. *Nature*, *75*(1949), 450–451, https://doi.org/10.1038/075450a0.

Giles, O. T., Romano, R., & Markkula, G. (2018). Bayesian analysis of subjective ranking data using Thurstonian Models: Tutorial, novel methods, and an open-source library. *PsyArXiv*, https://doi.org/10.31234/osf.io/t7szv.

Goddard, S. T. (1983). Ranking in tournaments and group decisionmaking. *Management Science*, *29*(12), 1384–1392, https://doi.org/10.1287/mnsc.29.12.1384.

Goldstein, D. G., McAfee, R. P., & Suri, S. (2014). The wisdom of smaller, smarter crowds. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14 (pp. 471—488). New York, NY, USA: Association for Computing Machinery.

Gordon, K. (1924). Group judgments in the field of lifted weights. *Journal of Experimental Psychology*, *7*(5), 398–400, https://doi.org/10.1037/h0074666.

Hamada, D., Nakayama, M., & Saiki, J. (2020). Wisdom of crowds and collective decision-making in a survival situation with complex information integration. *Cognitive Research: Principles and Implications*, *5*(1), 48, https://doi.org/10.1186/s41235-020-00248-z.

Han, Y. & Budescu, D. (2019). A universal method for evaluating the quality of aggregators. *Judgment and Decision Making*, *14*(4), 395–411, https://doi.org/10.1017/S1930297500006094.

Herrero, C. & Villar, A. (2021). Group decisions from individual rankings: The Borda–Condorcet rule. *European Journal of Operational Research*, *291*(2), 757–765, https://doi.org/10.1016/j.ejor.2020.09.043.

Herzog, S. M. & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–237, https://doi.org/10.1111/j.1467-9280.2009.02271.x.

Herzog, S. M. & Hertwig, R. (2014). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 218–232, https://doi.org/10.1037/a0034054.

Himmelstein, M., Budescu, D. V., & Ho, E. H. (2023). The wisdom of many in few: Finding individuals who are as wise as the crowd. *Journal of Experimental Psychology: General*, *152*(5), 1223–1244, https://doi.org/10.1037/xge0001340.

Hora, S. C. (2004). Probability judgments for continuous quantities: Linear combinations and calibration. *Management Science*, *50*(5), 597–604, https://doi.org/10.1287/mnsc.1040.0205.

Jenness, A. (1932). The role of discussion in changing opinion regarding a matter of fact. *The Journal of Abnormal and Social Psychology*, *27*(3), 279–296, https://doi.org/10.1037/h0074620.

Johnson, T. R. & Kuhn, K. M. (2013). Bayesian Thurstonian models for ranking data using JAGS. *Behavior Research Methods*, *45*, 857–872, https://doi.org/10.3758/s13428-012-0300-3.

Jose, V. R. R. & Winkler, R. L. (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting*, *24*(1), 163–169, https://doi.org/10.1016/j.ijforecast.2007.06.001.

Ju, J., Zhang, P., & Anderson, T. (2015). Project ranking using partial ranks. In *2015 Portland International Conference on Management of Engineering and Technology (PICMET)* (pp. 472–477).: IEEE.

Juni, M. Z. & Eckstein, M. P. (2017). The wisdom of crowds for visual search. *Proceedings of the National Academy of Sciences*, *114*(21), E4306–E4315, https://doi.org/10.1073/pnas.1610732114.

Kameda, T., Toyokawa, W., & Tindale, R. S. (2022). Information aggregation and collective intelligence beyond the wisdom of crowds. *Nature Reviews Psychology*, *1*(6), 345–357, https://doi.org/10.1038/s44159-022-00054-y.

Kao, A. B., et al. (2018). Counteracting estimation bias and social influence to improve the wisdom of crowds. *Journal of The Royal Society Interface*, *15*(141), 20180130, https://doi.org/10.1098/rsif.2018.0130.

Keck, S. & Tang, W. (2020). Enhancing the wisdom of the crowd with cognitive-process diversity: The benefits of aggregating intuitive and analytical judgments. *Psychological Science*, *31*(10), 1272–1282, https://doi.org/10.1177/0956797620941840.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–89, https://doi.org/10.2307/2332226.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, *77*(3), 217–273, https://doi.org/10.1016/0001-6918(91)90036-y.

Klugman, S. F. (1947). Group and individual judgments for anticipated events. *The Journal of Social Psychology*, *26*(1), 21–28, https://doi.org/10.1080/00224545.1947.9921728.

Knight, H. C. (1921). A comparison of the reliability of group and individual judgments. Master's thesis, Columbia University.

Krupinski, E. A. (2010). Current perspectives in medical image perception. *Attention, Perception, & Psychophysics*, *72*(5), 1205–1217, https://doi.org/10.3758/APP.72.5.1205.

Larrick, R. P., Burson, K. A., & Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, *102*(1), 76–94, https://doi.org/10.1016/j.obhdp.2006.10.002.

Larrick, R. P. & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127, https://doi.org/10.1287/mnsc.1050.0459.

Lee, M. D. (2024). Using cognitive models to improve the wisdom of the crowd. Manuscript submitted for publication.

Lee, M. D. & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, *9*(3), 258–272, https://doi.org/10.1017/S1930297500005799.

Lee, M. D., Danileiko, I., & Vi, J. (2018). Testing the ability of the surprisingly popular method to predict NFL games. *Judgment and Decision Making*, *13*(4), 322–333, https://doi.org/10.1017/S1930297500009207.

Lee, M. D. & Ke, M. Y. (2022). Modeling individual differences in beliefs and opinions using Thurstonian models. In J. Musolino, J. Sommer, & P. Hemmer (Eds.), *The cognitive science of belief: A multidisciplinary approach* (pp. 488–511). Cambridge University Press.

Lee, M. D., Liu, E., & Steyvers, M. (2015). The roles of knowledge and memory in generating top-10 lists. In D. C. Noelle & R. Dale (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1267–1272). Austin, TX: Cognitive Science Society.

Lee, M. D., Steyvers, M., de Young, M., & Miller, B. (2011a). A model-based approach to measuring expertise in ranking tasks. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1304–1309). Austin, TX: Cognitive Science Society.

Lee, M. D., Steyvers, M., de Young, M., & Miller, B. (2012). Inferring expertise in knowledge and prediction ranking tasks. *Topics in Cognitive Science*, *4*(1), 151–163, https://doi.org/10.1111/j.1756-8765.2011.01175.x.

Lee, M. D., Steyvers, M., & Miller, B. (2014). A cognitive model for aggregating people's rankings. *PloS One*, *9*(5), e96431, https://doi.org/10.1371/journal.pone.0096431.

Lee, M. D. & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*, 114–127, https://doi.org/10.3758/s13423-017-1238-3.

Lee, M. D. & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

Lee, M. D., Zhang, S., & Shi, J. (2011b). The wisdom of the crowd playing The Price Is Right. *Memory & Cognition*, *39*(5), 914–923, https://doi.org/10.3758/s13421-010-0059-7.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*(2), 149–188, https://doi.org/10.1006/obhd.1998.2804.

Li, X., Yi, D., & Liu, J. S. (2022). Bayesian analysis of rank data with covariates and heterogeneous rankers. *Statistical Science*, *37*(1), 1–23, https://doi.org/10.1214/20-STS818.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In H. Jungermann & G. De Zeeuw (Eds.), *Decision making and change in human affairs: Proceedings of the fifth research conference on subjective probability,* *utility, and decision making, Darmstadt, 1–4 September, 1975* (pp. 275–324). Springer Netherlands.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* chapter 22. Cambridge University Press.

Lin, A. Y., Huynh, A., Barrington, L., & Lanckriet, G. (2013). Search and discovery through human computation. In P. Michelucci (Ed.), *Handbook of Human Computation*. Springer New York.

Lin, A. Y., Huynh, A., Lanckriet, G., & Barrington, L. (2014). Crowdsourcing the unknown: The satellite search for Genghis Khan. *PloS ONE*, *9*(12), 1–17, `https://doi.org/10.1371/journal.pone.0114046`.

Lin, L. & Goodrich, M. A. (2010). A Bayesian approach to modeling lost person behaviors based on terrain features in Wilderness Search and Rescue. *Computational and Mathematical Organization Theory*, *16*, 300–323, `https://doi.org/10.1007/s10588-010-9066-2`.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*(6), 1231–1243, `https://doi.org/10.1037/0022-3514.47.6.1231`.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025, `https://doi.org/10.1073/pnas.1008636108`.

Lyon, A. & Pacuit, E. (2013). The wisdom of crowds: Methods of human judgement aggregation. In P. Michelucci (Ed.), *Handbook of Human Computation* (pp. 599–614). New York, NY: Springer.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299, `https://doi.org/10.1037/a0036677`.

Mayer, M. & Heck, D. W. (2023). Cultural consensus theory for two-dimensional location judgments. *Journal of Mathematical Psychology*, *113*, 102742, `https://doi.org/10.1016/j.jmp.2022.102742`.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10, `https://doi.org/10.1016/j.intell.2008.08.004`.

Miller, M. K., Wang, G., Kulkarni, S. R., Poor, H. V., & Osherson, D. N. (2012). Citizen forecasts of the 2008 US presidential election. *Politics & Policy*, *40*(6), 1019–1052, `https://doi.org/10.1111/j.1747-1346.2012.00394.x`.

Montgomery, L. E., Baldini, C. M., Vandekerckhove, J., & Lee, M. D. (2024). Where's Waldo, Ohio? Using cognitive models to improve the aggregation of spatial knowledge. *Computational Brain & Behavior*, *7*(2), 242–254, https://doi.org/10.1007/s42113-024-00200-0.

Montgomery, L. E. & Lee, M. D. (2022). The wisdom of the crowd and framing effects in spatial knowledge. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44. https://escholarship.org/uc/item/0h95m7m4.

Olsson, H. & Loveday, J. (2015). A comparison of small crowd selection methods. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the Thirty-Seventh Annual Meeting of the Cognitive Science Society* (pp. 1769–1774). Austin, TX: Cognitive Science Society. https://cognitivesciencesociety.org/wp-content/uploads/2019/03/cogsci15_proceedings.pdf.

Paese, P. W. & Sniezek, J. A. (1991). Influences on the appropriateness of confidence in judgment: Practice, effort, information, and decision-making. *Organizational Behavior and Human Decision Processes*, *48*(1), 100–130, https://doi.org/10.1016/0749-5978(91)90008-H.

Page, L. & Clemen, R. T. (2013). Do prediction markets produce well-calibrated probability forecasts? *The Economic Journal*, *123*(568), 491–513, https://doi.org/10.1111/j.1468-0297.2012.02561.x.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, volume 124 (pp. 1–10). Vienna, Austria. https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf.

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535, https://doi.org/10.1038/nature21054.

Prolific (2022). Online participant recruitment for surveys and market research. https://www.prolific.co.

Qing, S. & Fang, L. (2021). Research on the Intelligent Combat Decision-Making under the Simulation and Deduction System. In *2021 International Conference on Big Data and Intelligent Decision Making (BDIDM)* (pp. 206–209). Guilin, China: IEEE.

Ray, R. (2006). Prediction markets and the financial "wisdom of crowds". *The Journal of Behavioral Finance*, *7*(1), 2–4, https://doi.org/10.1207/s15427579jpfm0701_1.

Romney, A. K., Batchelder, W. H., & Weller, S. C. (1987). Recent applications of cultural consensus theory. *Americal Behavioral Scientist*, *31*(2), 163–177, https://doi.org/10.1177/000276487031002003.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*(2), 313–338, https://doi.org/10.1525/aa.1986.88.2.02a00020.

Ronis, D. L. & Yates, J. F. (1987). Components of probability judment accuracy: Individual consistency and effects of subject matter and assessment method. *Organizational Behavior and Human Decision Processes*, *40*, 193–218, https://doi.org/10.1016/0749-5978(87)90012-4.

Russo, J. E. & Schoemaker, P. J. H. (1992). Managing overconfidence. *Sloan Management Review*, *33*(2), 7–17.

Schvaneveldt, R. W., Durso, F. T., Goldsmith, T. E., Breen, T. J., Cooke, N. M., Tucker, R. G., & De Maio, J. C. (1985). Measuring the structure of expertise. *International Journal of Man-Machine Studies*, *23*(6), 699–728, https://doi.org/10.1016/s0020-7373(85)80064-x.

Selker, R., Lee, M. D., & Iyer, R. (2017). Thurstonian cognitive models for aggregating top-n lists. *Decision*, *4*(2), 87–101, https://doi.org/10.1037/dec0000056.

Shafir, E. (1993). Choosing versus rejecting: Why some options are both better and worse than others. *Memory & Cognition*, *21*(4), 546–556, https://doi.org/10.3758/BF03197186.

Shaw, M. E. (1932). A comparison of individuals and small groups in the rational solution of complex problems. *The American Journal of Psychology*, *44*(3), 491–504, https://doi.org/10.2307/1415351.

Simoiu, C., Sumanth, C., Mysore, A., & Goel, S. (2019). Studying the "wisdom of crowds" at scale. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*(1), 171–179, https://doi.org/10.1609/hcomp.v7i1.5271.

Steegen, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. (2014). Measuring the crowd within again: A pre-registered replication study. *Frontiers in Psychology*, *5*(786), 1–8, https://doi.org/10.3389/fpsyg.2014.00786.

Steyvers, M., Lee, M. D., Miller, B., & Hemmer, P. (2009). The wisdom of crowds in the recollection of order information. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22*, volume 22 (pp. 1785–1793). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2009/file/4c27cea8526af8cfee3be5e183ac9605-Paper.pdf.

Stock, J. H. & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*(6), 405–430, https://doi.org/10.1002/for.928.

Surowiecki, J. (2004). *The wisdom of crowds*. New York, NY: Doubleday, 1st edition.

Thomas, B., Coon, J., Westfall, H. A., & Lee, M. D. (2021). Model-based wisdom of the crowd for sequential decision-making tasks. *Cognitive Science*, *45*(7), e13011, https://doi.org/10.1111/cogs.13011.

Thurstone, L. L. (1927a). A law of comparative judgement. *Psychological Review*, *34*, 273–286.

Thurstone, L. L. (1927b). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, *21*(4), 384–400, https://doi.org/10.1037/h0065439.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, *95*, 261–289, https://doi.org/10.1007/s10994-013-5401-4.

Tversky, A. & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458, https://doi.org/10.1126/science.7455683.

van Doorn, J., Westfall, H. A., & Lee, M. D. (2021). Using the weighted Kendall distance to analyze rank data in psychology. *The Quantitative Methods for Psychology*, *17*(2), 154–165, https://doi.org/10.20982/tqmp.17.2.p154.

Vul, E. & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647, https://doi.org/10.1111/j.1467-9280.2008.02136.x.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, *39*(2), 176–190, https://doi.org/10.1287/mnsc.39.2.176.

Welsh, M. B. & Begg, S. H. (2018). More-or-less elicitation (MOLE): Reducing bias in range estimation and forecasting. *EURO Journal on Decision Processes*, *6*(1-2), 171–212, https://doi.org/10.1007/s40070-018-0084-5.

Wysokiński, M., Marcjan, R., & Dajda, J. (2014). Decision support software for search & rescue operations. *Procedia Computer Science*, *35*, 776–785, https://doi.org/10.1016/j.procs.2014.08.160.

Yaniv, I. & Schul, Y. (1997). Elimination and inclusion procedures in judgment. *Journal of Behavioral Decision Making*, *10*(3), 211–220, https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<211::AID-BDM250>3.0.CO;2-J.

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science*, *36*, 452–470, https://doi.org/10.1111/j.1551-6709.2011.01223.x.

Zhang, S. & Lee, M. D. (2010). Cognitive models and the wisdom of crowds: A case study using the bandit problem. In *Proceedings of the annual meeting of the cognitive science society*, volume 32 (pp. 1118–1123). https://escholarship.org/uc/item/5k71f0vd.