Computational Brain & Behavior (2024) 7:242–254
https://doi.org/10.1007/s42113-024-00200-0

ORIGINAL PAPER

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

Lauren E. Montgomery[1] · Charles M. Baldini[1] · Joachim Vandekerckhove[1] · Michael D. Lee[1]

Accepted: 18 February 2024 / Published online: 8 April 2024

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 and Heck (2023) in which participants provided point estimates for the locations of European cities.

Keywords Wisdom of the crowd · Spatial knowledge · Expertise · Cognitive modeling

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 and Hertwig, 2009; Keck and 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).

Parts of this research were presented at the 2022 Annual Joint Meeting of the Society for Mathematical Psychology and the International Conference on Cognitive Modeling, the Meeting of the European Mathematical Psychology Group in 2022, and the 58th Edwards Bayesian Research Conference held in 2023.

✉ Lauren E. Montgomery
lmontgo1@uci.edu

[1] Department of Cognitive Sciences, University of California Irvine, Irvine, CA 92697-5100, USA

One way to address the challenges of multidimensional behavior and structured expertise is to use cognitive models (see Lee, 2024 submitted, 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., 2011), 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 and 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 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 and 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 and 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.

Fig. 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

Experimental Design

Experimental Interface

A screenshot of the experimental interface is shown in Fig. 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.

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.

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 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 the 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.

Behavioral Analyses

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 in 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 the 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, Fig. 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.

Crowd Performance

We used the arithmetic mean and a weighted arithmetic mean as *statistical* wisdom of the crowd estimates. The simple wis-

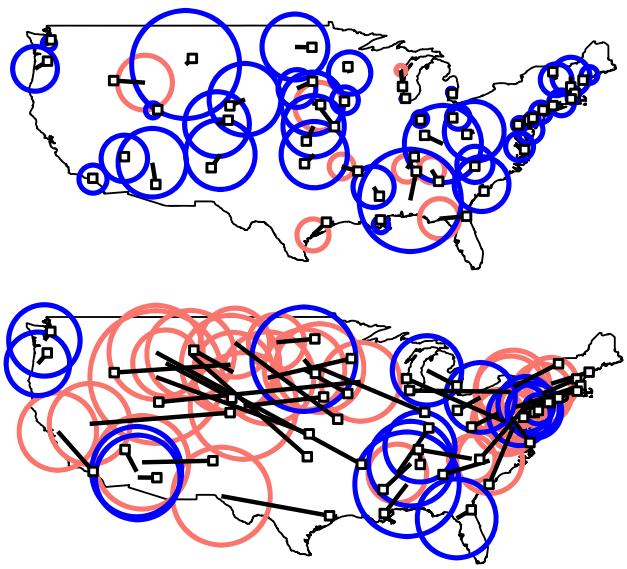

Fig. 2 The true locations of the 48 city locations compared with the estimated locations for a well-performed (top) and poorly-performed (bottom) participant. The true locations of the cities are shown by squares, and the error is shown by the line connecting the true location to the participant's point 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

dom 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 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. 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, Fig. 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 Fig. 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.

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.

Model of Point Estimates

Our approach to modeling the point estimate uses several key features of the cognitive model developed by Mayer and 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

$$\boldsymbol{y}_{ij} \sim \text{bivariate Gaussian}\big(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{ij}\big), \tag{1}$$

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:

$$\mu_{j1} \sim \text{uniform}\big(0, 2.44\big) \tag{2}$$
$$\mu_{j2} \sim \text{uniform}\big(0, 1\big). \tag{3}$$

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.

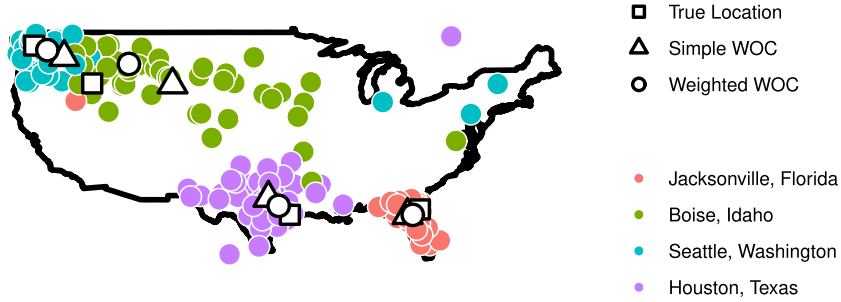

Fig. 3 The 50 participants' estimates for four cities: Jacksonville (coral), Boise (green), Seattle (teal), and Houston (lilac). The city's true location is shown as a square, the simple wisdom of the crowd estimate is shown as a triangle, and the weighted wisdom of the crowd estimate is shown as a circle

The covariance matrix $\mathbf{\Sigma}_{ij}$ in Eq. 1 is specified as

$$\mathbf{\Sigma_{ij}} = \begin{bmatrix} \lambda_{j1}^2 + \sigma_i^2 + \beta_{ij}^2 & \rho_j \sqrt{\lambda_{j1}^2 + \sigma_i^2 + \beta_{ij}^2} \sqrt{\lambda_{j2}^2 + \sigma_i^2 + \beta_{ij}^2} \\ \rho_j \sqrt{\lambda_{j1}^2 + \sigma_i^2 + \beta_{ij}^2} \sqrt{\lambda_{j2}^2 + \sigma_i^2 + \beta_{ij}^2} & \lambda_{j2}^2 + \sigma_i^2 + \beta_{ij}^2 \end{bmatrix}, \tag{4}$$

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:

$$\beta_{ij} \sim \text{Gaussian}_+\left(0, \frac{1}{\omega_i^2}\right). \tag{5}$$

The model developed by Mayer and 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:

$$\sigma_i \sim \text{uniform}\left(0, 1\right) \tag{6}$$
$$\omega_i \sim \text{uniform}\left(0, 1\right). \tag{7}$$

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 Fig. 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:

$$\lambda_{j1} \sim \text{Gaussian}_+\left(\mu_{\lambda_1}, 1/\sigma_{\lambda_1}^2\right) \tag{8}$$
$$\lambda_{j2} \sim \text{Gaussian}_+\left(\mu_{\lambda_2}, 1/\sigma_{\lambda_2}^2\right) \tag{9}$$
$$\mu_{\lambda_1}, \mu_{\lambda_2} \sim \text{uniform}\left(0, 2\right) \tag{10}$$
$$\sigma_{\lambda_1}, \sigma_{\lambda_2} \sim \text{uniform}\left(0, 1\right). \tag{11}$$

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

$$\rho_j \sim \text{uniform}\left(-1, 1\right). \tag{12}$$

Model of Radius Judgments

Mayer and 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_{ij}^r 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 Eq. 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

$$y_{ij}^r \sim \text{Gaussian}\left(\alpha_i \sqrt{\max(\boldsymbol{\lambda}_j)^2 + \sigma_i^2 + \beta_{ij}^2}, 1/\tau^2\right). \tag{13}$$

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:

$$\alpha_i \sim \text{uniform}\left(0, \sqrt{2.44^2 + 1^2}\right) \tag{14}$$
$$\tau \sim \text{uniform}\left(0, 1\right). \tag{15}$$

Model Identifiability

The full cognitive model defined by Eqs. 1–15 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.

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 Eqs. 4 and 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 Eqs. 5 and 7 and replace σ_i in Eq. 6 with σ. The assumption of no individual-by-city expertise requires removing Eqs. 5 and 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 Eqs. 10–11 and adjust Eqs. 8–9. Specifically, λ_{j1} and λ_{j2} in Eqs. 8–9 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 Eqs. 13, 14, and 15. This specific form of the model can be applied to data involving only point estimates, such as those collected by Mayer and Heck (2023).

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 are available in the supplementary information.

Results

Performance Results

Each of the 12 models makes predictions 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 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 Fig. 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 errors of the individuals are shown as a yellow

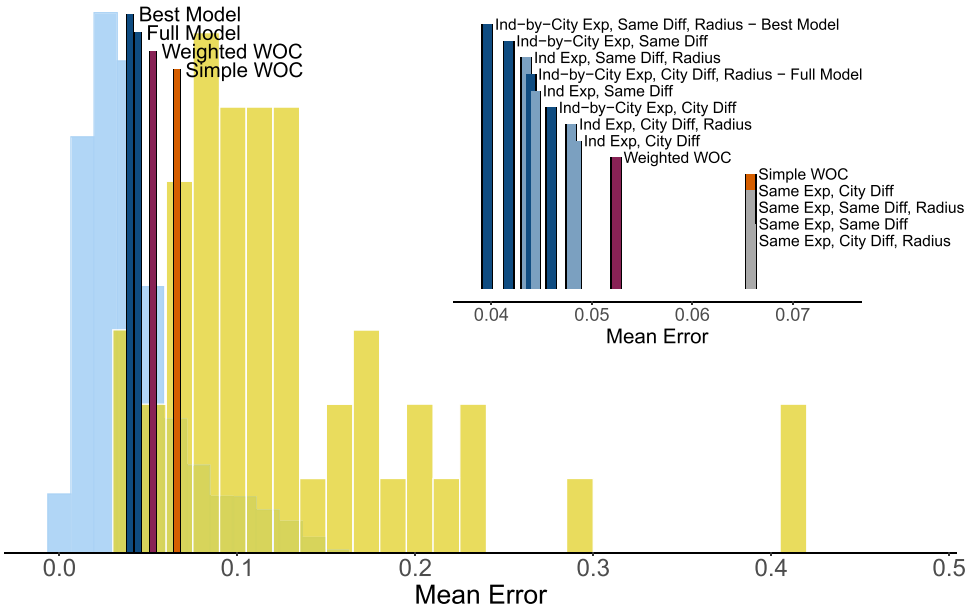

Fig. 4 The main panel shows the distribution of individual mean error in yellow and the mean error of statistical and model-based crowd aggregates by vertical lines. The posterior distribution for the best-performing model's mean error is shown in blue. The vertical bars in the inset panel provide a magnified view of the performance of model-based and statistical estimates, with color coding to indicate the assumption each model makes about expertise

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 Fig. 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 the 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.

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 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 Fig. 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 aver-

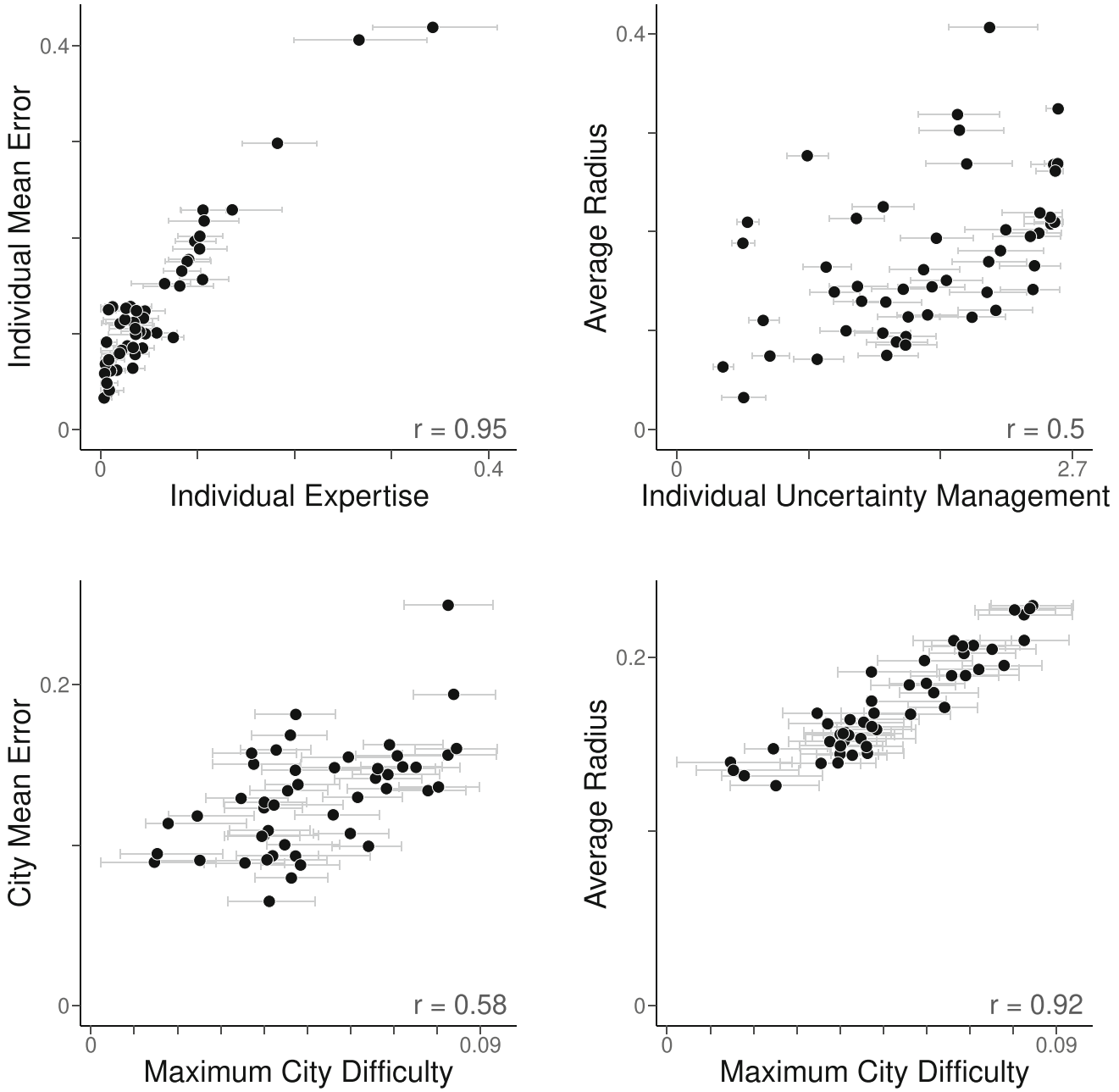

Fig. 5 The relationship between parameter values and behavioral measures of individual differences in terms of expertise and uncertainty, and city differences in terms of difficulty. See main text for details

age 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 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 Fig. 5 show that the key model parameters of expertise, uncertainty management, and city difficulty cor-

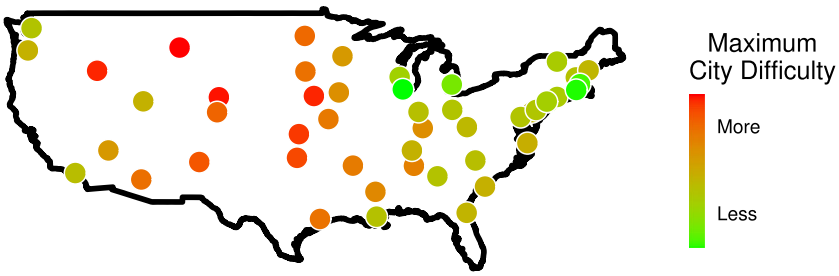

Fig. 6 A visualization of inferred maximum city difficulty. Circles are located on the true locations of the cities. Cities inferred to be less difficult are in bright green, and cities inferred to be more difficult are in bright red

relate well with conceptually related behavioral measures. Figure 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 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.

Application to Mayer and Heck (2023)

To evaluate the replicability and generalizability of our findings, we applied the same cognitive models to the data set collected by Mayer and Heck (2023). Mayer and 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 and 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 and 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 and 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

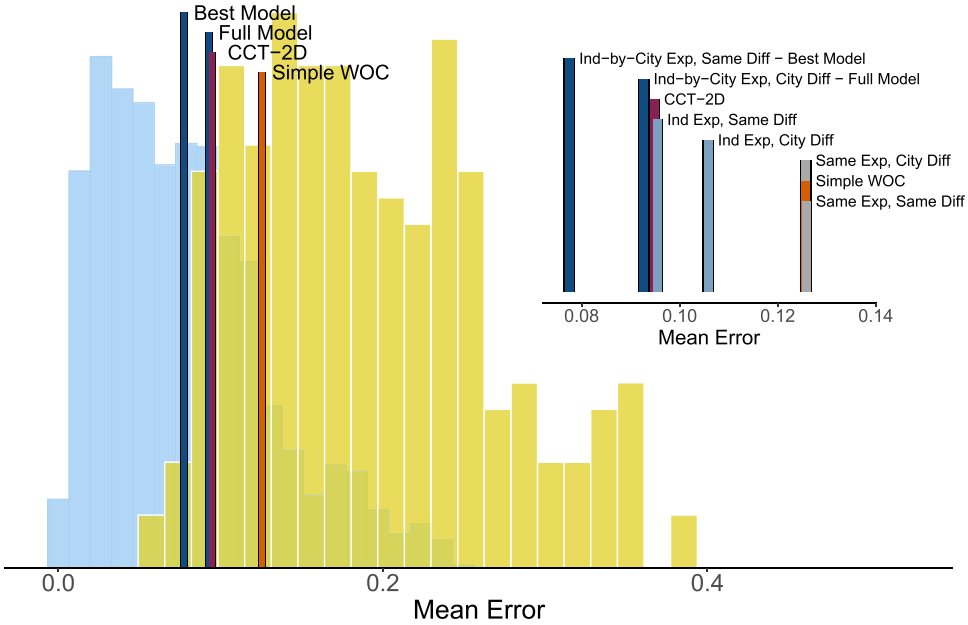

Fig. 7 Analysis of Mayer and Heck's (2023) data. The main panel shows the distribution of individual mean error in yellow and the mean error of statistical and model-based estimates by vertical lines. The posterior distribution for the best-performing model's mean error is shown in blue. The vertical 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

approach. Because of its CCT-2D foundations, there are a few differences in the details of Mayer and 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 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 and Heck (2023) model is shown in Fig. 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 and 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 and Heck (2023) model performed slightly better than our model that made the same psychological assumptions, presumably due to their different priors.

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 submitted) 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 instruc-

tions 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 and 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 possible on 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.

Acknowledgements We thank the Bayesian Cognitive Modeling lab at the University of California, Irvine, for their feedback, and Daniel Heck, Brandon Turner, and an anonymous reviewer for useful comments.

Author Contribution LEM and MDL conceived and designed the study. LEM, CMB, and MDL designed the experiment, and CMB programmed the experiment. LEM collected the experimental data. LEM, MDL, and JV developed the model. LEM performed the modeling analysis. CMB wrote the first draft of the Stimuli and Procedure subsections. LEM wrote the rest of the first draft. LEM and MDL contributed to revisions of the paper. All authors read and approved the final manuscript.

Funding This research was supported by the US Air Force Research Laboratory's Continuous Learning Branch. LEM's collaboration was enabled through an appointment to the Oak Ridge Institute for Science and Education (ORISE) Summer Research Internship Program. MDL's collaboration was enabled through an appointment to the ORISE Faculty Research Program. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the US Air Force, Department of Defense, or the US Government.

Data and Code Availability The US cities' data set is available on an Open Science Framework project page at https://www.osf.io/ve8t9/. The Mayer and Heck (2023) data set is available on an Open Science Framework project page at https://www.osf.io/jbzk7/. The JAGS code, code for any analyses presented in this paper, and other supplementary materials are available on an Open Science Framework project page at https://www.osf.io/ve8t9/.

Declarations

Ethics Approval This project was approved via exempt self-determination by the University of California Irvine (UCI) Institutional Review Board (IRB). This project also made use of de-identified archival data that was obtained from https://www.osf.io/jbzk7/.

Consent to Participate Informed consent was obtained from all individuals who participated in the study.

Consent for Publication Not applicable

Competing Interests The authors declare no competing interests.

References

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

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

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

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

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

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. https://doi.org/10.1093/actrade/9780198796206.001.0001

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. (2002). 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. (2002). 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

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

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. https://doi.org/10.1145/2600057.2602886

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

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

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

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.

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

Lee, M. D. (2024 submitted). 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., Steyvers, M., & Miller, B. (2014). A cognitive model for aggregating people's rankings. *PLoS ONE, 9*(5). https://doi.org/10.1371/journal.pone.0096431

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., & Wagenmakers, E.-J. (2014). Bayesian cognitive modeling: A practical course. *Cambridge University Press*. https://doi.org/10.1017/cbo9781139087759

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

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. https://doi.org/10.1007/978-94-010-1276-8_19

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

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). Springer, New York. https://doi.org/10.1007/978-1-4614-8806-4_47

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

Montgomery, L. & 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*, vol. 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). Cognitive Science Society, Austin, TX. https://cognitivesciencesociety.org/wp-content/uploads/2019/03/cogsci15_proceedings.pdf

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)*, vol. 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

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 judgment 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

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

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

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

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

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). https://doi.org/10.1111/cogs.13011

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

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

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

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