Judgment and Decision Making, Vol. 9, No. 3, May 2014, pp. 259–273

Using cognitive models to combine probability estimates

Michael D. Lee* Irina Danileiko*

Abstract

We demonstrate the usefulness of cognitive models for combining human estimates of probabilities in two experiments. The first experiment involves people's estimates of probabilities for general knowledge questions such as "What percentage of the world's population speaks English as a first language?" The second experiment involves people's estimates of probabilities in football (soccer) games, such as "What is the probability a team leading 1–0 at half time will win the game?", with ground truths based on analysis of large corpus of games played in the past decade. In both experiments, we collect people's probability estimates, and develop a cognitive model of the estimation process, including assumptions about the calibration of probabilities and individual differences. We show that the cognitive model approach outperforms standard statistical aggregation methods like the mean and the median for both experiments and, unlike most previous related work, is able to make good predictions in a fully unsupervised setting. We also show that the parameters inferred as part of the cognitive modeling, involving calibration and expertise, provide useful measures of the cognitive characteristics of individuals. We argue that the cognitive approach has the advantage of aggregating over latent human knowledge rather than observed estimates, and emphasize that it can be applied in predictive settings where answers are not yet available.

Keywords: probability estimation, wisdom of the crowd, cognitive modeling, graphical models, individual differences.

1 Introduction

The wisdom of the crowd effect involves combining the decisions or estimates made by a group of people, and observing that the resulting group decision or estimate has a good level of performance relative to that of the individuals. The effect is based on the amplification of the common knowledge shared by people, and the aggregation of the different knowledge dispersed over people (Surowiecki, 2004). Often the mechanisms by which knowledge is combined are simple statistical ones, such as taking means, medians, or modes.

A different approach is to treat the challenge of aggregation as a cognitive modeling problem (Lee, Zhang, & Shi, 2011; Merkle & Steyvers, 2011; Turner, Steyvers, Merkle, Budescu, & Wallsten, 2013). The basic data that need to be combined are behavioral observations, generated by cognitive decision-making processes based on people's knowledge. The motivation for a cognitive approach is that it is the knowledge people have, and not their behavioral estimates, that should be combined. This view recognizes that people can be prone to biases and distortions in how they represent and express information. A good cognitive model of their representations and processes can serve to "undo" the distortion, and allow for useful inferences about the knowledge people have. Combining this inferred knowledge can potentially lead to group answers that outperform the statistical combination of the observed behavioral estimates.

A natural framework for approaching wisdom of the crowds aggregation as a cognitive modeling problem is provided by hierarchical Bayesian models (Lee, 2011; Lee & Wagenmakers, 2013). Hierarchical models are able to represent knowledge at different levels of abstraction, and so can represent both individual- and group-level information in latent parameters. Using Bayesian inference, these parameters can be linked to observed behavioral data through models of decision-making processes.

Turner et al. (2013) use hierarchical Bayesian methods to pursue the problem of using individual judgments to forecast probabilistic events. Their models incorporate a key insight from the existing literature on human estimation of probabilities, which is that people may be miscalibrated in their perception of probabilities (see Brenner, Kohler, Liberman, & Tversky 1996; Lichtenstein, Fischoff, & Phillips, 1982; Yates, 1990, for reviews). The models developed by Turner et al. (2013) explicitly incorporate calibration processes, and build on the work of Budescu and Johnson (2011) and Merkle (2010) to use hierarchical methods to allow for individual differences in calibration.

A second insight from the existing literature is that there are individual differences in expertise (Weiss & Shanteau, 2014). This aspect of individual differences is not in-

We thank Jon Baron and two anonymous reviewers, whose comments significantly improved this paper. Support from UCI UROP Grant 88405 and SURP Fellowship 56930 is gratefully acknowledged.

*Department of Cognitive Sciences, 3151 SSPA Zot 5100, University of California Irvine, Irvine CA, USA, 92697. Email: mdlee@uci.edu; idanilei@uci.edu.

https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

cluded in the models developed by Turner et al. (2013). One way they can be included within a hierarchical modeling approach is developed by Lee, Steyvers, de Young, and Miller (2012), in the context of different but related wisdom of the crowd problem involving ranking data. The basic idea is to assume that people's representations are all centered on a common ground truth, but the expertise of the individual determines how precisely they represent the truth.

In this paper we develop a hierarchical model for the cognitive aggregation of individual behavior to the problem of combining human estimates of probabilities. Our model differs in three important ways from previous models that forecast binary events based on combining people's probability estimates (e.g., Turner et al., 2013; Baron, Mellers, Tetlock, Stone, & Ungar, 2014; Satopää, Baron, Foster, Mellers, Tetlock, & Ungar, 2014). First, our model includes not just calibration processes, but also allows for individual differences in both calibration and the expertise of individuals. Second, we evaluate the model by collecting two new data sets in which people are asked to estimate directly the probabilities of events. The first data set involves general knowledge questions, and the second involves questions coming from the real-world statistical environment provided by football (soccer) games. The detail provided by continuous ground truth probabilities, as opposed to binary outcomes generated from those probabilities, allows for more detailed model evaluation. Thirdly, our modeling approach is completely unsupervised, in the sense that it never receives feedback about true probabilities (nor outcomes of probabilistic events generated from those probabilities). We find that our cognitive model for combining people's estimates outperforms simple statistical methods, and that there is interpretable structure about how individual performed in the inferred individual differences within the model.

2 Experiments

We conducted two experiments to collect people's estimates of probabilities. The first experiment involved general knowledge questions, and the second experiment involved questions relating to football (soccer) games. Because the experiments are methodologically extremely similar, we describe both together. The full datasets are provided as supplementary materials along with this paper on the page for this issue: http://journal.sjdm.org/vol9.3.html.

2.1 Participants

For each experiment, 145 participants were recruited using Amazon Mechanical Turk. Participants were paid US$1 for completing the questionnaire within the Qualtrics survey software interface. Completing an experiment took an average of about 20 minutes.

2.2 Probability estimation questions

2.2.1 General knowledge questions

We constructed 40 questions requiring the estimation of a probability or a percentage, as detailed in Table 1. The answers were found from a variety of sources including the 2013 CIA World Factbook, Government websites, the websites of the relevant professional societies, and Wikipedia. The questions were presented in a random order for each participant. After the questions had been answered, each participant was asked a final question "On a scale of 1 (very poor) to 7 (very well), how well do you think you estimated probabilities?"

Because participants were recruited through Amazon Mechanical Turk they were not supervised and had access to the internet while completing the estimation task. To address the possibility that participants could search for answers, we vetted questions to insure they could not be immediately answered through a simple Google search. This meant that a search using the question text or keywords from the question did not display the answer in the top matches returned by Google visible on the returned page from the search. Of course, participants could have conducted more detailed searches to find the answers, but we could find no evidence for this behavior in the accuracies of their answers or the time taken to provide them.

2.2.2 Football questions

Sports like football provide real-world statistical environments that have been widely analyzed (see, Albert, Bennett, & Cochran, 2005, for a relatively recent anthology of papers). Because most people have some level of understanding of a popular sport like football, and statistics characterizing the outcomes of games are readily available, it is a convenient setting for studying the psychology of probability estimation (e.g., Bar-Hillel, Budescu, & Amar, 2008). To compile the necessary statistical characterization of the football environment, the details of 6072 first-division professional games played between 2001 and 2011 in the domestic leagues of a large number of countries were obtained from http://soccerbot.com. From the information available in these records, we parsed the sequence of goals scored by the home and away team. For example, the information recorded for a US Major League Soccer game between home team Chicago Fire and and away team Los Angeles Galaxy was that the home team scored goals in the 1st and 84th minutes, and the away team scored a goal in the 78th minute.

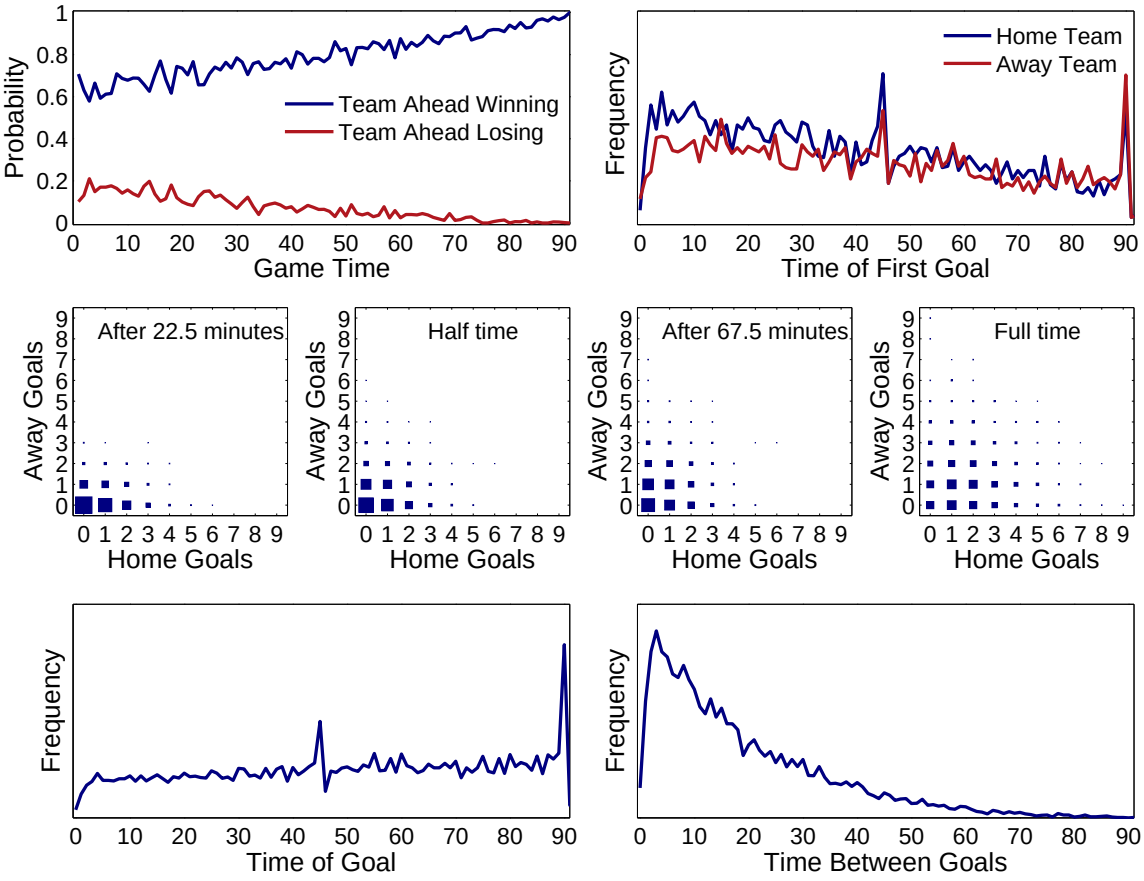
https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

Table 1: The 40 general knowledge questions and their answers.

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

https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

Figure 1: Analyses of the football game environment, based on 6072 games from first-division games played in domestic leagues between 2001 and 2011.



From these goal scoring data, a variety of statistical analyses of the football game environment are possible. Figure 1 shows a set of analyses on which our probability estimation task questions are based. The panel in the top left shows the (frequentist) probability of the team currently ahead eventually winning or losing a game, as a function of the time at which they are currently ahead. The panel in the top right shows the probability a team will score their first goal at a certain time, for both home and away teams. The sequence of panels in the middle row show the distribution of game scores (i.e., home team goals and away team goals) at four different times during a game. The bottom panels show the distributions of the times goals are scored, and the length of time between goals.

The 40 estimation questions are detailed in Table 2, together with the empirical ground truth found by analyses of the game data shown in Figure 1. The questions were developed in terms of eight types — such as questions about probabilities of the team ahead winning — with five specific questions for each type. The question types were always completed in the same order listed in Table 2, but the order of the specific questions within each type was randomized for each participant. In addition, the questionnaire began by asking participants to self-rate their football expertise on a seven-point scale, and answer seven multiple-choice trivia questions involving football facts.

2.3 Basic results

Figure 2 summarizes the performance of the individuals in both probability estimation tasks. Performance is measured as the mean absolute difference between a participant's estimates and the answers over all 40 questions. The histograms of stick figures show the distribution of performance for all of the participants in the general knowledge (upper panel) and football (lower panel) experiments. It is clear that there is a wide range of performance across people, with the best-performed participants within about 0.1 of the true probabilities on average, and the worst-performed participants 0.3 or 0.4 from the truth on average. Inset with each histogram in Figure 2 are two panels

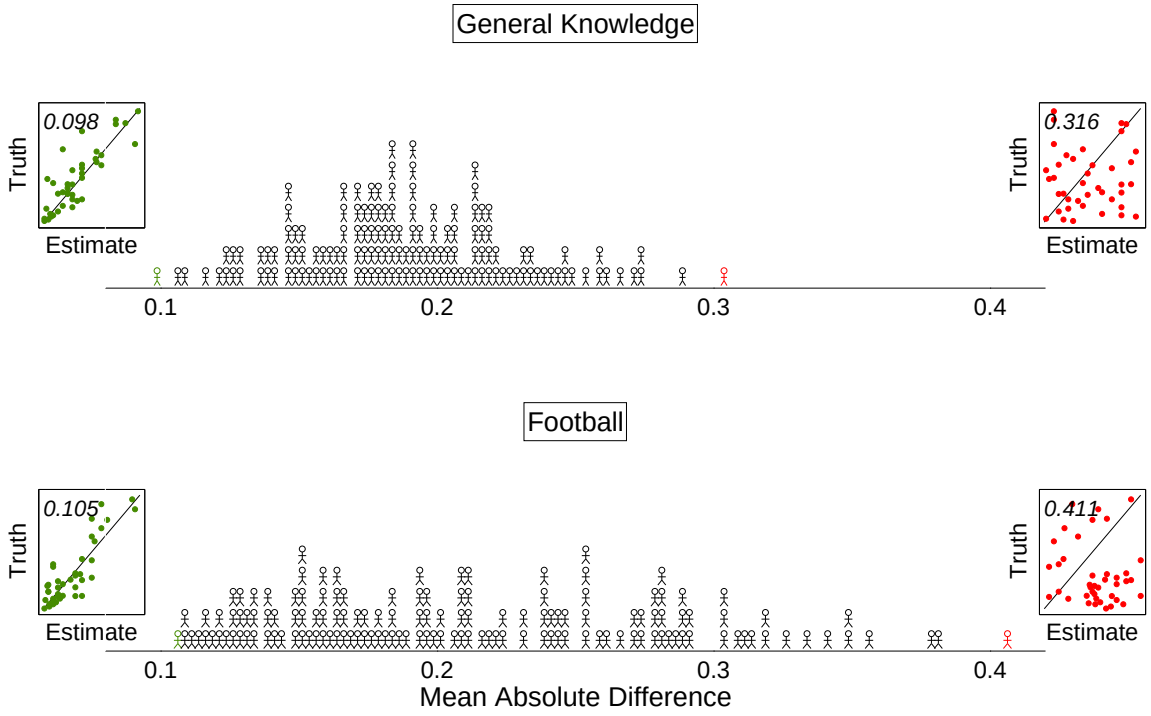
https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

Table 2: The 40 football estimation questions and their empirical answers.

Question	Answer
Team Ahead Winning	
What is the probability that a team that is ahead at the 5th minute will win?	0.59
What is the probability that a team that is ahead at the 25th minute will win?	0.71
If a team is ahead at half-time, what is the probability that this team will win?	0.78
If a team is leading at the 60th minute, what is the probability that it will win?	0.88
If a team is leading at the 85th minute, what is the probability that it will win?	0.96
Games at Half Time	
What percentage of games are 0–1 at half-time, with the away team leading?	15%
What percentage of games are 2–0 at half-time, with the home team leading?	8%
A score of 2–1 at half-time, with the home team leading, occurs in what percentage of games?	4%
A score of 3–2 at half-time, with the home team leading, occurs in what percentage of games?	0%
1–0 scores at half-time, with the away team leading, occur in what percentage of games?	20%
Team Ahead Losing	
What is the probability that a team that is ahead at the 15th minute will lose?	0.30
If a team is ahead at the 35th minute, what is the probability that this team will lose?	0.19
What is the probability that a team that is ahead at half-time will lose?	0.29
If a team is leading at the 50th minute, what is the probability that it will lose?	0.26
If a team is leading at the 80th minute, what is the probability that it will lose?	0.37
Games at Full Time	
What percentage of games are 1–0 at full-time, with the home team winning?	11%
What percentage of games are 0–2 at full-time, with the away team winning?	4%
A score of 2–1 at full-time, with the home team winning, occurs in what percentage of games?	9%
A score of 1–2 at full-time, with the away team winning, occurs in what percentage of games?	6%
2–3 scores at full-time, with the away team winning, occur in what percentage of games?	2%
Goals Scored	
What percentage of goals are scored between the 0 and 15th (inclusive) minute?	13%
What percentage of goals are scored during the first half of the game?	44%
In between the 40th and 55th minute, what percentage of goals are scored?	19%
In between the 55th and 75th minute, what percentage of goals are scored?	24%
Goals scored between the 85th and 90th minute make up what percentage of total goals?	11%
Games With Draws	
What percentage of games are at draws (ties) at half-time?	43%
What percentage of games are at a draw (tie) at full-time?	25%
What percentage of games are at a 0–0 draw (tie) at half-time?	31%
What percentage of games are at a 2–2 draw (tie) at full-time?	5%
What percentage of games are at a 1–1 draw (tie) at full-time?	11%
First Goals Scored	
What percentage of first goals of a game are scored between the 0 and 10th minute?	15%
What percentage of first goals of a game are scored between the 15th and 35th minute?	27%
In between the 45th and the 65th minute, what percentage of the first goals of a game are scored?	22%
In between the 70th and the 85th minute, what percentage of the first goals of a game are scored?	12%
What percentage of goals are first goals scored between the 80th and 90th minute?	9%
Another Goal	
After a goal is scored, what is the probability that a goal is scored in the following 5 minutes?	0.21
After a goal is scored, what is the probability that a goal is scored in the following 20 minutes?	0.64
After a goal is scored, what is the probability that a goal is scored in the following 30 minutes?	0.79
What is the probability that 45 minutes separates two goals of either team?	0.92
What is the probability that 10 minutes separates two goals of either team?	0.39

https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

Figure 2: Histograms of stick people showing the distribution of performance, measured as the mean absolute difference between estimates and true probabilities, for all participants in both the general knowledge (upper) and football (lower) experiments. The inset panels show, for each experiment, the relationship between the estimates and the answers for the best- and worst-performed participants.



showing the performance of the best- and worst-performed participants. These panels show scatter plots of the relationship between the estimates provided by these participants to all of the questions, and the true answers.

3 A model for combining probability estimates

3.1 Theoretical assumptions

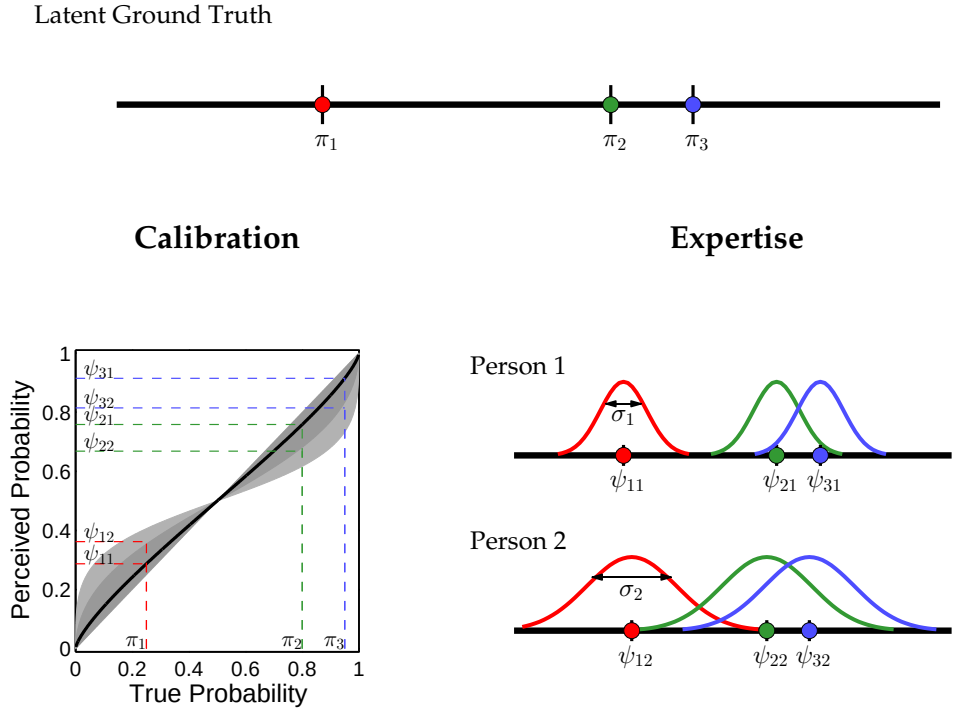
The primary data collected from each of the experiments consist of probability (percentage) estimates of the 145 participants to all 40 questions. It is straightforward, for each question, to find the mean and median estimate, as standard statistical approaches to combining people's estimates.

Developing a cognitive model of the data requires making assumptions about how people represent probabilities, and how they produce estimates. Figure 3 shows the basic assumptions and motivations for the cognitive model we developed and applied to the data. The founding assumption is that the true probability for each question is a latent parameter, represented by π_i for the the ith question. The goal of a cognitive modeling approach is to specify how that knowledge is represented within individuals, and how decision processes act on the knowledge to produce the observed data.

The first psychological assumption involves the miscalibration of probabilities. It has often been found in probability estimation tasks that people systematically over-estimate small probabilities and under-estimate large probabilities (see, Zhang & Maloney, 2012, for a review). This means that the behavioral estimates generated by people are distorted versions of a person's latent knowledge of the probability. Building a calibration process into a model allows for the distortion to be corrected. The goal is to combine people's latent knowledge, free from the effects of miscalibration. One simple calibration model is shown in the bottom-left panel of Figure 3 and involves a non-linear function that maps true to perceived probabilities, consistent with the over-estimating small probabilities and over-estimating large ones. A number of different mathematical forms, motivated in part by different theoretical assumptions, have been proposed for this function, although they share the same basic qualitative properties (Cavagnaro, Pitt, Gonzalez, & Myung, 2013, Goldstein & Einhorn, 1987; Gonzalez & Wu, 1999; Prelec, 1998;

https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

Figure 3: The theoretical framework for our cognitive model of probability estimation. The ith probability is assumed to have a latent truth π_i that is subjected to calibration and expertise processes in producing an observed estimate. Calibration operates according to a non-linear function that maps true to perceived probabilities, such that small probabilities are over-estimated and large probabilities are under-estimated. Expertise controls how precisely a perceived probability is reported through the standard deviation of the Gaussian distribution from which the behavioral estimate is sampled. Both the level of calibration and expertise processes are controlled by participant-specific parameters that allow for individual differences.



Turner et al., 2013; Tversky & Kahneman, 1992; Zhang & Maloney, 2012).

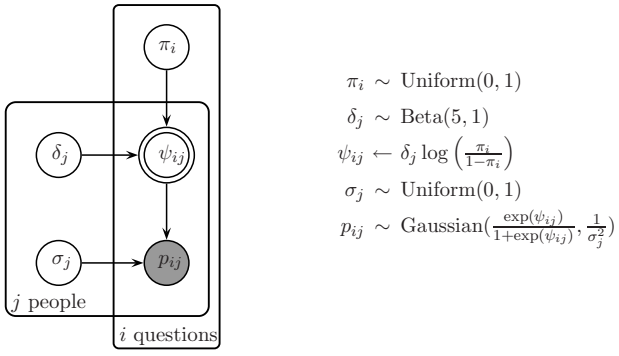
We chose to use a linear-in-log-odds functional form with a single parameter capturing the magnitude of over- and under-estimation, because it has a natural interpretation that helps in defining the model. The single parameter δ_j for the jth participant scales the log-odds $\log(\pi_i/(1-\pi_i))$ representation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

This calibration function is shown in the bottom-left of Figure 3. The expected (mean) prior transformation is shown by the solid line and the 90% and 99% credible regions are shown by successive shading. The transformations of three different true probabilities are shown by three lines, which trace the ith true probability π_i to the perception of that probability by the jth person ψ_{ij}. In the specific examples shown, the first person is well calibrated, so ψ_{11}, ψ_{21}, and ψ_{31} are very similar to π_1, π_2, and π_3. The second person, however, is miscalibrated, and so their perceived probability ψ_{12} overestimates the small true probability π_1, while the perceived probabilities ψ_{22} and ψ_{32} underestimate the large true probabilities π_2 and π_3.

The second psychological assumption made by the model involves expertise. Different people seem likely to have different levels of understanding of the true environmental probabilities, and this will affect the precision of their knowledge and the accuracy of their answers. One way to incorporate this assumption, used successfully in a related modeling problem by (Lee et al., 2012), is to assume people's estimates are draws from Gaussian distributions that have a level of variability associated with their knowledge. This approach is shown at the bottom-right of Figure 3, When the jth participant answers the ith

https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

Figure 4: Graphical model for behavioral estimates of probabilities made by a number of participants for a number of questions. The latent true probability π_i for the ith question is calibrated according to a parameter δ_j for the jth participant to become the value ψ_{ij}. This calibrated values then produces an observed estimate p_{ij} according to the expertise σ_j of the participant.



question, the assumption is that their estimate comes from a Gaussian distribution that is centered on their perceived probability ψ_{ij}, but has a standard deviation σ_j. The assumption is that σ_j is a property of the participant, and is the same for all of the questions. In this way, the parameter σ_j represents the level of knowledge or expertise of the jth participant, with smaller values corresponding to greater expertise.

3.2 Graphical model

The graphical model in Figure 4 formalizes our cognitive model. Graphical models are a standard tool in statistics and machine learning (Jordan, 2004; Koller, Friedman, Getoor, & Taskar, 2007), and are becoming an increasingly popular approach for implementing and evaluating probabilistic models of cognitive processes (Lee, 2011; Lee & Wagenmakers, 2013; Shiffrin, Lee, Kim, & Wagenmakers, 2008). In graphical models, nodes represent variables and data, and the graph structure is used to indicate dependencies between variables. Continuous variables are represented with circular nodes and discrete variables are represented with square nodes. Observed variables, which are usually data or properties of an experimental design, are shaded and unobserved variables, which are usually model parameters, are not shaded. Plates are square boundaries that enclose subsets of the graph that have independent replications in the model. The attraction of graphical models is that they provide an interpretable and powerful language for expressing probabilistic models of cognitive processes, and can easily be analyzed using modern computational Bayesian methods. In particular, they can be implemented and evaluated in standard software like WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) and JAGS (Plummer, 2003) that automatically approximates the full joint posterior distribution of a model and data.

In Figure 4, the underlying latent probability π_i for the ith question is an unobserved and continuous variable, and so is shown as an unshaded circular node. These are the "true" answers to the probability questions that we want to infer. The behavioral data take the form of probability estimates p_{ij} given by the jth person for the ith question. These are observed continuous values, and so are shown as shaded circular nodes. The cognitive model describes the process that generates the observed behavior from the assumed latent knowledge. It is important to understand that the model is never provided with the answers to the questions. This means that the latent parameters inferred — the latent ground truths of the questions, and the calibration and expertise parameters of participants — are based solely on using the model to account for the generation of the behavioral data.

The graphical model naturally shows how the two core psychological assumptions convert the latent true probability to the observed behavioral estimate. First, the latent probability π_i is transformed according to the calibration function. Since ψ_{ij} is a function of π_i and δ_j it is shown as a double-bordered deterministic node. The extent of over- and under-estimation is controlled by the prior distribution of δ_j, which is naturally expressed as a beta distribution. As Figure 4 shows, we chose $\delta_j \sim \text{Beta}(5,1)$ because it gives most weight to large values of δ_j that will not transform the latent probabilities drastically, consistent with existing empirical findings and theory. We settled on the exact beta distribution by inspection of the prior distribution for the calibration function it defines, as shown in the bottom-left of Figure 3. It is important to note that we defined this prior, as we developed the model, *before* we used the model to analyze data and did not adjust the prior to optimize the results obtained.

The second processing stage produces the estimate p_{ij} as a draw from a Gaussian distribution. The mean is the calibrated probability ψ_{ij} re-expressed on the probability rather than log-odds scale as $\exp(\psi_{ij}) / (1 + \exp(\psi_{ij}))$. The standard deviation of the Gaussian distribution is σ_j for the jth person, and is given a simple weakly informative prior $\sigma_j \sim \text{Uniform}(0,1)$ (Gelman, 2006).[1] The plates in the graphical model in Figure 4 replicate over the questions and over the participants. The latent ground truth π_i for each question interacts with the calibration δ_j and expertise σ_j of participants to produce the observed data p_{ij}.

[1] Note that we follow the convention used by JAGS of parametrizing the Gaussian distribution in terms of mean and precision parameters.

https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

This model is related to the "hierarchical calibrate then average" graphical model presented by Turner et al. (2013), but there are important differences. The Turner et al. (2013) model accounts for the binary *outcomes* of probabilistic events (e.g., "whether a soccer team actually won a game"), whereas our model accounts for the underlying probabilities themselves (e.g., the latent probability the team will win the game). Of course, as part of predicting outcomes the Turner et al. (2013) model determines probabilities that could be assessed against the data from our experiments, and so the difference might be regarded as relatively superficial. But, it remains the case that these are not the data the model was designed to predict.

More fundamentally, the Turner et al. (2013) model does not incorporate individual differences in the representation of the true probabilities, and uses a different two-parameter form of the linear-in-log-odds calibration function, including an intercept parameter. This difference in modeling assumptions can probably be traced to the third — and most fundamental — difference between the two models. The Turner et al. (2013) model *observes* the outcomes of the binary events it is designed to predict, and relies on cross-validation methods for evaluation. Our modeling approach, in contrast, never presents the ground truth probabilities to the model. In machine learning terms, the modeling is fully unsupervised, and so models can be directly assessed in terms of their predictions, since there is no possibility of a model being able to over-fit data because of its complexity. This difference requires our model to specify *a priori* psychologically plausible distributions over models parameters, since they cannot be inferred from data, and so makes the model a more complete attempt to describe the processes involving in people's knowledge of probabilities, their estimation processes, and individual differences in both (Vanpaemel & Lee, 2012).

We also considered reduced versions of our model that included only the calibration or only the expertise assumption. This was done by maintaining either the calibration or the expertise elements of the graphical model in Figure 4, but not both, so that only one of the theoretical assumptions was incorporated in the model. Formally, the model that has only calibration used a single σ parameter for all participants, so that $p_{ij} \sim \text{Gaussian}\big(\exp(\psi_{ij}) / (1+\exp(\psi_{ij})), 1/\sigma^2\big)$, while the model that uses only individual differences has no calibration function, so that $p_{ij} \sim \text{Gaussian}\big(\pi_i, 1/\sigma_j^2\big)$. Considering these reduced models allows us to explore whether both calibration and individual differences are useful assumptions, and whether each makes a contribution above and beyond what the other provides.

4 Modeling results

We implemented the graphical model in Figure 4 using JAGS, and applied it to both the general knowledge and football probability estimation data sets. For both analyses we collected eight independent Markov chain Monte Carlo chains, each with 2000 burn-in samples that were discarded and 2000 collected samples. Standard measures of convergence and auto-correlation, including the $\hat{R}$ statistic (Gelman, 1996), were evaluated to validate the samples as good approximations to the posterior distribution. We implemented and analyzed the reduced models incorporating only calibration or individual differences in exactly the same way.

4.1 Estimation accuracy

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

The performance of each of these five measures — three based on cognitive models, and two on statistical summaries — is shown for the general knowledge experiment in Figure 5. The bottom panel shows the distribution of individual participant performance presented in Figure 2 and superimposes as vertical lines the performance of the five methods. The best performing method is the cognitive model with calibration and expertise, which produces estimates on average 0.125 from different from the true probabilities. The median of participant's answers is 0.127 different on average, followed by the reduced models assuming only calibration or expertise, which are 0.131 different on average. The mean of participant's answers is the worst-performed method, with an average difference of 0.135.

The inserted panels in Figure 5 show as scatter plots the relationship between the true answer and the answer generated by each method. The methods themselves are ordered from left to right from best performing to worst performing, as measured by the average difference between the true answer and the method's answer. It is clear that all of the wisdom of crowds methods perform relatively well, in relation to individual performance, with lower average differences than all but a few individuals.

Figure 6 provides the same analysis of estimation accuracy for the football questions. The model that includes calibration and expertise performs much better than the other approaches, being an average of 0.128 from the true empirical probabilities, and is again better performed than

https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

Figure 5: The performance of three cognitive models and two statistical methods in estimating probabilities for the general knowledge questions, and the relationship of their levels of performance to individual participants. The cognitive models assume calibration and expertise ("Calibrate+Expertise"), just calibration ("Calibrate") or just expertise ("Expertise"). The statistical methods are the median and the mean of individual responses for each question. The top panels show the relationship between true and estimated answers for all 40 questions for each method. The bottom panel shows the distribution of individual performance as stick figures and the levels of model performance as broken lines. The performance of the models and individuals is measured as mean absolute difference from true answers.

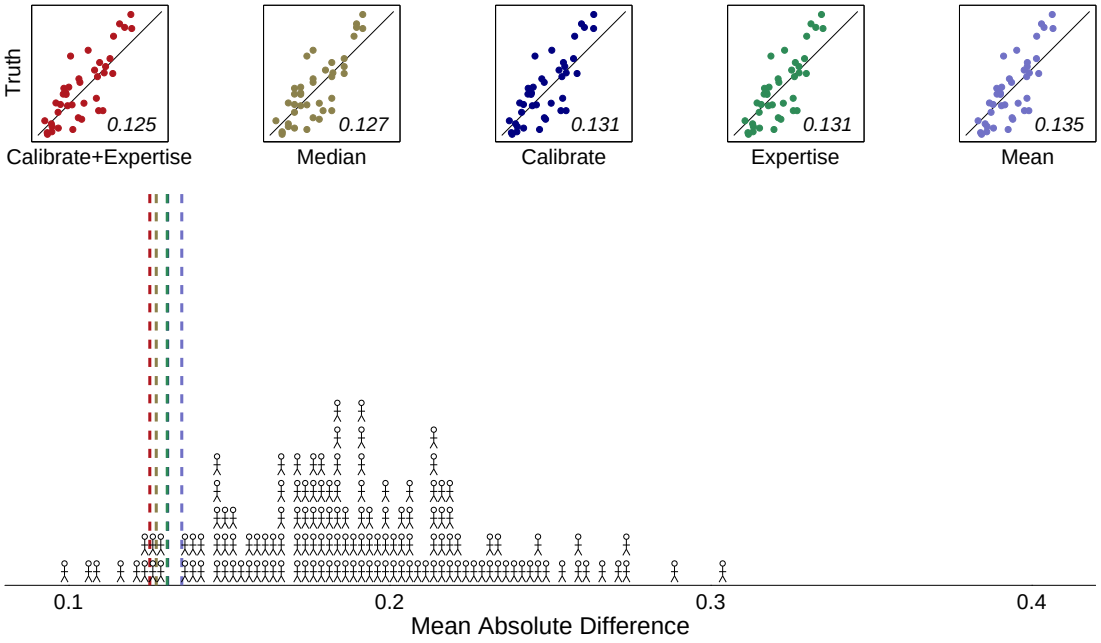
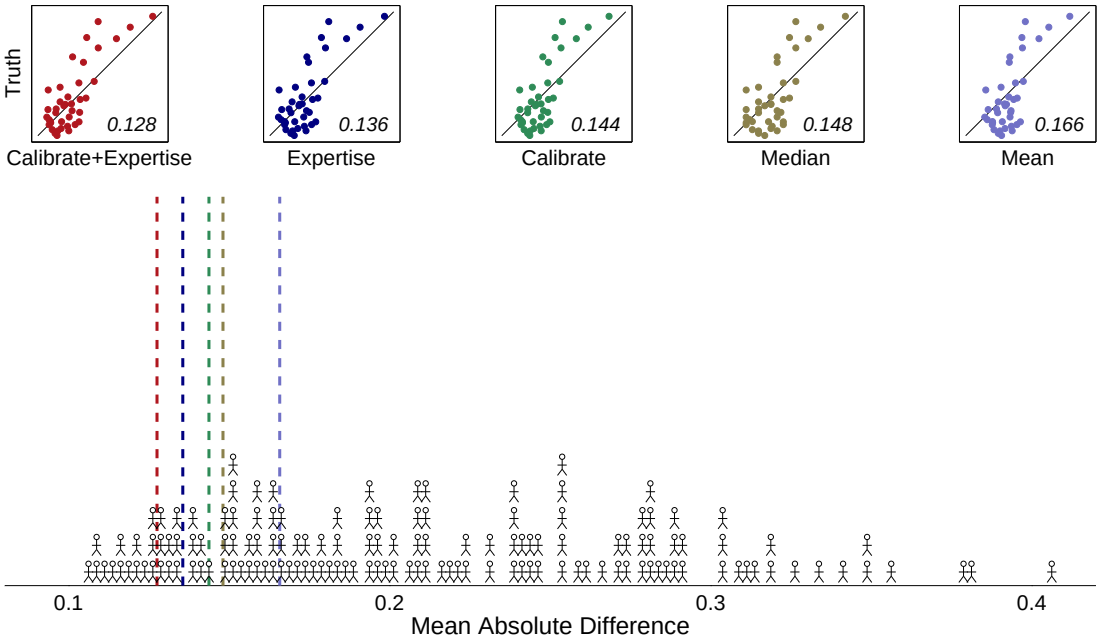


Figure 6: The performance of three cognitive models and two statistical methods in estimating probabilities for the football questions, and the relationship of their levels of performance to individual participants. The same information is presented in the same format as for the general knowledge questions presented in Figure 5.



https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

all but a few individual participants. The two reduced models also outperform both of the statistical approaches. Once again, the best-performed cognitive modeling aggregation methods are among the best-performed participants.

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

4.2 Expertise and calibration

The two parameters inferred for each participant are the σ_j measure of expertise and the δ_j measure of calibration. Their expected posterior values for each participant are shown in Figure 7 as scatter plots for both the general knowledge (left) and football (right) experiments. In both experiments a wide range of values are inferred for both parameters. The expertise parameter — which is a standard deviation for an assumed Gaussian distribution lying on the probability scale from 0 to 1 — ranges from about 0.1 to about 0.3. The calibration parameter — which is a multiple of log-odds — ranges from about 0.95 down to about 0.5 for the general knowledge experiment and lower to 0.3 or 0.4 for the football experiment. Thus, it seems clear that both parameters capture variation underlying the probability estimates produced by different participants. There is also no obvious strong correlation between the two parameters for either experiment, suggesting they capture, at least in part, different aspects of the variation in people's performance.

Also shown for the general knowledge experiment in Figure 7 are the detailed performance of four individual participants. These participants were selected at the "extremes" of the joint parameter space, to give an indication of the type of estimates inferred to have high and low expertise and strongly or weakly miscalibrated. The participant labeled "A" in the top-left panel is relatively accurate in their estimates and shows no systematic miscalibration, and is inferred to have high expertise and be well calibrated. Participant "B" in the bottom-left panel systematically over-estimates small probabilities but under-estimates large ones, and is inferred to be similarly expert but miscalibrated. Participants "C" and "D" show poor performance, and are inferred to be much less expert. Participant "C" does not appear to mis-estimate systematically while the participant "D" does over-estimate small probabilities often. The calibration parameters that the model infers are consistent with this difference.

4.3 Inferred expertise and performance

Figure 8 presents an analysis of how the inferred expertise of participants relates to their actual performance on the probability estimation tasks. The top-left panel shows their relationship for the general knowledge questions, while the bottom-left panel shows the relationship for the football questions. For both sets of questions there is a strong positive correlation, with people who were inferred to have greater expertise having performed better.

Figure 8 also shows how the various self-reported measures of expertise collected in the two experiments relate to performance. The top-right panel shows the relationship between self-reported expertise and performance in the general knowledge experiment. The two panels in the bottom row shows the relationship between self-reported football expertise and the number of trivia questions answered correctly in the football experiment. None of these self-reported measures are strongly correlated with performance.

5 Discussion

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

As part of formalizing the cognitive process people are assumed to use to make probability estimates, the cognitive model introduces parameters that control individual differences between people. In our model, one process related to calibration and another to expertise, and both involved a single parameter. One perspective on these processes and parameters is that they support the improved inference of the underlying probabilities. Inferring that an individual participant is miscalibrated allows that distortion to be "undone" in the inference of the latent probabilities. Inferring that an individual participant is relatively

https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

Figure 7: The expected posterior expertise σ_j and calibration δ_j parameters for each participant in the general knowledge (left) and football (right) experiments. For the general knowledge experiment, four participants are highlighted and the scatter plot of their estimates relative to the answers are shown in inserted panels.

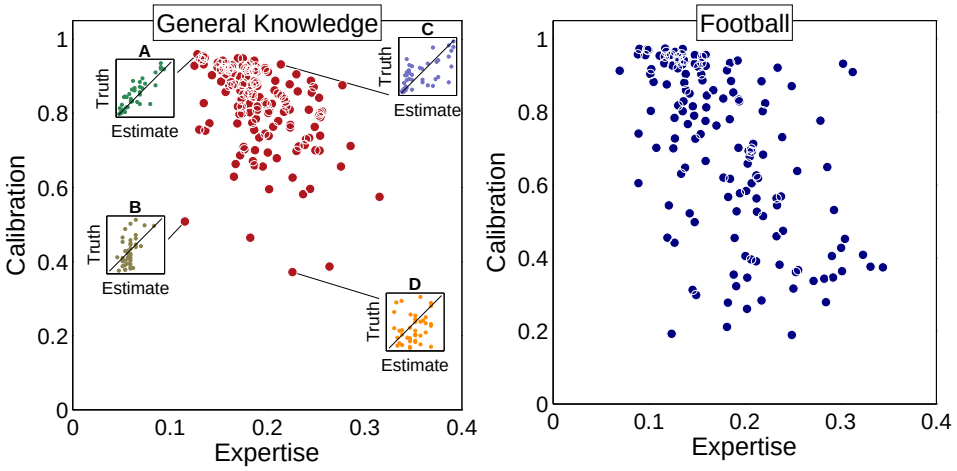
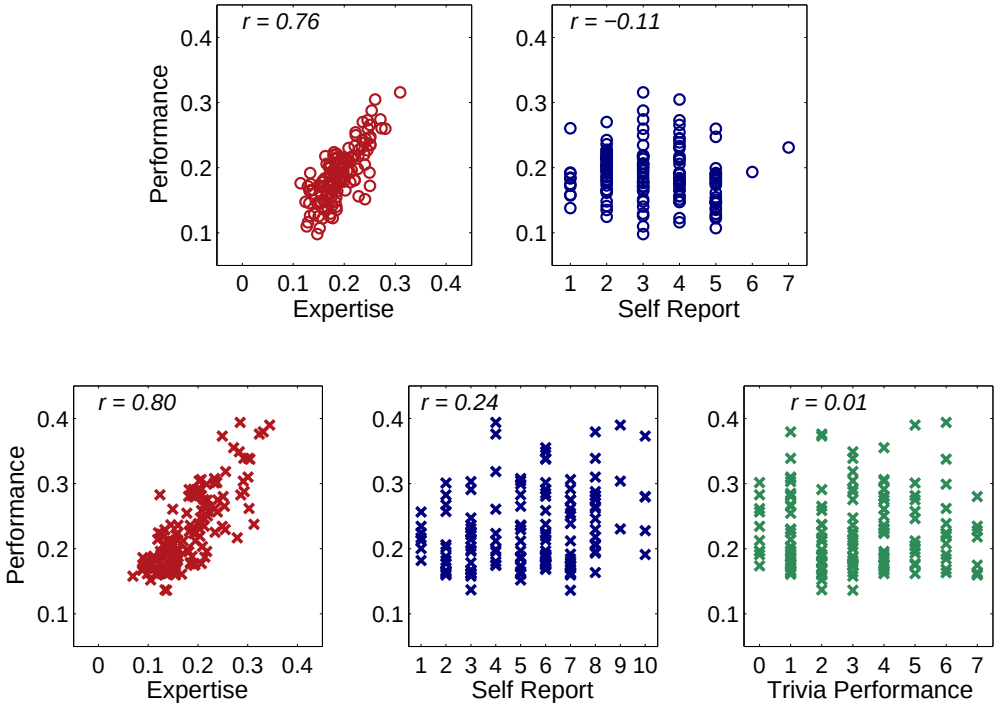


Figure 8: The relationship between model-based inferences of individual expertise, self reported measures of expertise, and actual performance in estimating probabilities across individuals. The top two panels relate to the general knowledge questions, and show how the model-based expertise and self reported expertise correlate with performance in estimating probabilities. The bottom three panels relate to the football questions, and show how the model-based expertise, self reported expertise, and trivia question performance relate to performance in estimating probabilities. For each scatter plot the Pearson correlation coefficient is also shown.



more expert allows their estimates to be "up-weighted" in inferring the latent probabilities. From this machine learning or statistical perspective (Turner et al., 2013), the graphical model in Figure 4 can be conceived as a method for the non-linear averaging of behavioral estimates of probabilities that performs well in approximating the underlying true probabilities.

A complementary cognitive science perspective is that

https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

the effectiveness of the modeling provides evidence for its core assumptions as important components of human decision-making. From this perspective, the behavioral data collected in the experiments provide further empirical evidence for the systematic miscalibration of probability estimates and for the presence of significant individual differences in expertise. The analyses of the parameters associated with these processes, presented in Figures 7 and 8, provide insight into basic psychological characteristics of the people providing estimates. The inferred parameter values identify people who are calibrated or miscalibrated and relatively more or less expert.

We emphasize that these inferences are made *without* knowledge of the answers to the questions. The graphical model in Figure 4 does not contain the answers to the questions. This is an especially important point in understanding the contribution the success of our model makes to assessing psychological theory. Leaving the data to be predicted unobserved forces not just the data generating process — involving calibration and individual differences — to be specified in the model, but also the values of the parameters that control those processes. It is not possible, for example, to infer appropriate values for a calibration function from the performance of the model on previously-made predictions. Instead, the priors for these sorts of parameters must formalize the relevant psychological assumptions, making the model more theoretically complete, and more readily falsifiable (Vanpaemel & Lee, 2012). This means, for example, it is not a trivial result that the model was able to infer the miscalibration of participant "B" in Figure 7. It is clear from the scatter plot that this participant over-estimates small probabilities and under-estimates large probabilities systematically, but the model was able to determine this miscalibration *without* reference to the answers.

Similarly, the ability to infer expertise without reference to the answers means our modeling approach makes *predictions* about the performance of the individuals. This is because inferred value of the σ_j expertise parameters is available before the answers to the questions are considered. Thus, the strong positive correlation between expertise and accuracy shown in Figure 8 has obvious applied possibilities, especially since the basic self-report measures we considered do not correlate with performance in the same way. In our experiments, of course, the answers were determined as the questions were generated. But the same modeling approach would apply in genuinely predictive settings for which answers are not yet known. For example, people could be asked to estimate probabilities for football games for the upcoming season, so that answers are only available after the season has finished. The model we have developed would immediately make predictions about people's individual expertise and our current results suggest those predictions could be usefully accurate.

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

Of course, the two experiments presented here constitute only a limited test of the model. Each experiment is a significant empirical undertaking — especially the football experiment for which answers had to be determined through a time-consuming compilation and analysis of a large data-set — but additional experiments in additional domains should be conducted to test modeling performance further. It would be particularly worthwhile to undertake a genuinely predictive test, asking questions for which answers cannot yet be known. We note that existing work using cognitive models for aggregating human judgments in probabilistic forecasting is structured slightly differently from our experiments. In the standard predictions tasks in this area, people are asked whether or not a real-world event, such as "Obama will win the 2014 US Presidential election" will happen. These probability estimates are amenable to cognitive aggregation, but are challenging to assess because only a binary ground truth is observed (e.g., Turner & Steyvers, 2011). The fact that Obama won the 2012 US Presidential election gives relatively little information for distinguishing between people who gave probability estimates of 0.7, 0.8 or 0.9 for this event. In contrast, our statistical characterization of a real-world environment like football games allows empirical ground truth for event probabilities to be measured.

https://doi.org/10.1017/S1930297500005799 Published online by Cambridge University Press

It is also true that the graphical model in Figure 4 is not the only possible formalization of the psychological assumptions about calibration and individual differences. Our model assumed that true probabilities were first calibrated, and then subject to individual differences in how accurately they led to behavioral estimates. It would be possible to develop models in which there were first individual differences in the perception of latent true probabilities, and then calibration before estimation. It would also be possible to relax the assumption that each individual has exactly one calibration curve, as parametrized by δ_j, and one level of knowledge, as parametrized by σ_j, that apply to all questions. And it would also be possible to introduce different and more general calibration functions, or allows for mixtures of different latent grounds truths, or accommodate a host of more sophisticated psychological assumptions about how people estimate probabilities. All sort of theoretical assumptions are naturally implemented within the graphical modeling framework, using hierarchical extensions and latent mixtures to formalize the processes that generate behavior from knowledge (Lee, 2011).

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

References

Albert, J., Bennett, J., & Cochran, J. J. (Eds.). (2005). *Anthology of statistics in sports.* SIAM.

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

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014) Two Reasons to Make Aggregated Probability Forecasts More Extreme. *Decision Analysis* Published online in Articles in Advance 19 Mar 2014. urlhttp://dx.doi.org/10.1287/deca.2014.0293.

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

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

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

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

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

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

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

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

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning.* Cambridge, MA: MIT Press.

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

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

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course.* Cambridge, England: Cambridge University Press.

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

Lichtenstein, S., Fischoff, 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* (pp.

306–334). Cambridge, England: Cambridge University Press.

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

Merkle, E. C. (2010). Calibrating subjective probabilities using hierarchical Bayesian models. In *Social Computing, Behavioral Modeling, and Prediction (SBP) 2010* (Vol. 6007, p. 13-22).

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

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.* Vienna, Austria.

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

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., & Ungar, L. H. (2014). Combining multiple probability predictions using a simple logit model. *International Journal of Forecasting, 30*(2), 344–356.

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

Surowiecki, J. (2004). *The wisdom of crowds.* New York: Random House.

Turner, B. M., & Steyvers, M. (2011). A wisdom of the crowd approach to forecasting. In *2nd NIPS workshop on Computational Social Science and the Wisdom of Crowds.*

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (in press). Forecast aggregation via recalibration. *Machine Learning.*

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

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

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

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

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