Steyvers, M., Miller, B. (2015). Cognition and Collective Intelligence. In T.W. Malone and M.S. Bernstein (Eds.) Handbook of Collective Intelligence. MIT Press, pp. 119-138

Cognition and Collective Intelligence

Mark Steyvers and Brent Miller

Cognitive and psychological research provides useful theoretical perspectives for understanding what is happening inside the mind of an individual in tasks such as memory recall, judgment, decision making, and problem solving—including meta-cognitive tasks, in which individuals reflect on their own performance or that of others. Although certain types of interactions between group members can allow groups to collectively process information (e.g., transactive memory; see Wegner 1987) or to utilize shared mental states through patterns in the environment (Norman 1993), the focus of this chapter will be on cognitive processes contained wholly within single minds that can affect group behavior.

We humans differ from many other collectively intelligent organisms in groups (see the chapter on Human–Computer Interaction) in that we are measurably intelligent independent of one another. Understanding the cognitive processes within individuals can help us understand under what conditions collective intelligence might form for a group and how we might optimize that group's collective performance. These components, alone or in concert, can be understood to form the basic building blocks of group collective intelligence.

Consider the classic estimation task in which a group of individuals must determine the number of marbles in a jar. In the simplest conceptualization of this task, each individual independently provides an estimate and a statistical average of the estimates is taken as the crowd's answer. The statistical aggregate over individuals can often lead to an answer that is better than that arrived at by most of the individuals. This has come to be known as the "wisdom of crowds" effect (Ariely et al. 2000; Davis-Stober et al. 2014; Surowiecki 2004; Wallsten et al. 1997). Given simple, idealized tasks, it would appear that extracting the collective intelligence from a group of individuals merely requires choosing a

suitable statistical aggregation procedure—no psychology or understanding of the underlying cognitive processes is necessary. However, if we start to make more realistic assumptions about the estimation task or change it to make it more like complex, real-world situations, it quickly becomes obvious how psychological factors can come into play. Suppose, for example, that individuals give judgments that are systematically biased (e.g., they may overestimate the number of marbles in a jar because the marbles differ in size). How can we know what the potential biases are, and how to correct for them? Suppose that some individuals are better at a task than others, or do not understand the task, or aren't even paying attention. How do we identify the judgments that are more accurate? What are the measures that we can use to identify experts? If individuals share information about their judgments and their reasoning, how does the sharing affect the results? To fully understand how collective intelligence arises from a group of individuals, and how a group's collective wisdom can be improved, it is necessary to consider what is going on inside the human mind.

In this chapter, we will review the cognitive and psychological research related to collective intelligence. We will begin by exploring how cognitive biases can affect collective behavior, both in individuals and in groups. Next we will discuss expertise and consider how more knowledgeable individuals may behave differently and how they can be identified. We will also review some recent research on consensus-based models and meta-cognitive models that identify knowledgeable individuals in the absence of any ground truth. We will then look at how the sharing of information by individuals affects the collective performance, and review a number of studies that manipulate how information is shared. Finally, we will look at collective intelligence within a single mind.

Identifying and Correcting for Biases

Whether or not a group collectively arrives at sensible judgments depends largely on whether the individuals are sensible. The literature provides many examples of cognitive biases that can systematically distort the judgment of individuals (see, e.g., Hogarth 1975; Kahneman et al. 1982). For example, human probabilistic judgments may be overconfident about reported probabilities, may neglect the event's base rate, or may be biased by the desirability of the outcomes (Kahneman and Tversky 2000; Gilovich et al. 2002; Massey et al. 2011). Biased

misperceptions of likelihood can have deleterious effects on entire economies (Taleb 2007). Conversely, individuals may often be sensitive to extraneous information that can be irrelevant to the judgment task at hand (Goldstein and Gigerenzer 2002). Systematic distortions that affect individuals' judgments can also affect group performance. Although uncorrelated errors at the level of individual judgments can be expected to average out in the group, systematic biases and distortions cannot be averaged out by using standard statistical averaging approaches (Simmons et al. 2011; Steyvers et al. 2014).

It has been shown to be possible to get subjects to debias their own estimates, at least to a degree, by training individuals in the potential biases of estimation (Mellers et al. 2014). Alternatively, it is possible, by understanding what these cognitive biases are, to correct them before performing statistical aggregation. In some domains, such as predicting the likelihood of low-probability events, subjects are systematically overconfident (Christensen-Szalanski and Bushyhead 1981). In judging other events that occur more frequently, as in weather forecasting, experts have more opportunity to properly calibrate their responses (Wallsten and Budescu 1983). When expert judgments are tracked over a period of time, it is possible to learn and correct for systematic biases. Turner et al. (2014) used hierarchical Bayesian models to learn a recalibration function for each forecaster. The calibrated individual estimates were then combined using traditional statistical methods, and the resulting aggregation was found to be more accurate than aggregates of non-calibrated judgments. Satopää et al. (2014) have proposed similar recalibration methods that shift the final group estimates, using either weighted or unweighted aggregation of the individual responses.

Human judgment can also be error-prone and inconsistent when information between interrelated events needs to be connected. For example, when people judge the likelihood of events that are dependent on one another, the result can lead to incoherent probability judgments that do not follow the rules of probability theory (Wang et al. 2011). Probabilities for interrelated events are coherent when they satisfy the axioms of probability theory. For example, the probability of a conjunction of events (A and B) has to be equal to or less than the probability of the individual events (A or B). However, people may not always connect these interrelations in logical ways and may fail to produce coherent probability judgments. Failure of coherence can occur at the individual level (Mandel 2005), but also can occur at the aggregate level in prediction markets (Lee et al. 2009). Similarly,

probability judgments that are incoherent at the individual level cannot be expected to become coherent by averaging across individuals (Wang et al. 2011). Incoherence may persist even in the presence of financial incentives (Lee et al. 2009). Wang et al. (2011) proposed a weighted coherentization approach that combines credibility weighting with coherentization so that the aggregate judgments are guaranteed to obey the rules of probability; for instance, when they asked participants to forecast the outcome of the 2008 U.S. presidential election, some of the questions were about elementary events but others involved negations, conditionals, disjunctions, and conjunctions (e.g., "What is the probability that Obama wins Vermont and McCain wins Texas?"). Sometimes humans make errors in estimation when the task environment encourages them to do so. In a competitive environment with information sharing, there may be an advantage to not giving one's best estimates to others. Ottaviani and Sørensen (2006) studied professional financial forecasters and found that the incentive to distinguish oneself from one's fellow forecasters outweighed the traditional goal of minimizing estimation error. Depending on the nature of the competition, fairly complex cognitive strategies can be employed to generate answers that are not representative of individuals' true estimates. On the televised game show *The Price is Right*, contestants bid in sequential order on the price of an item, and the winner is the contestant who comes closest to the price without exceeding it. Contestants often give estimates that are quite far below the actual price (and presumably quite far from what they belief that price to be) in order to increase their odds of winning. Aggregation approaches that model the strategic considerations of such competitive environments and attempt to aggregate over inferred beliefs outperform standard aggregation methods (Lee et al. 2011). When competition is employed, a winner-take-all format with minimal information may be best suited to get the most useful estimates from individuals; there is reason to believe that people will be more likely to employ any unique information they may have to make riskier but more informative estimates for aggregation (Lichtendahl et al. 2013).

Identifying Expert Judgments

The presence of experts in a group can have a significant effect on the accuracy and behavior of collective intelligence. The ability to identify and use these experts is an important application in a wide range of real-world settings. Society expects experts to provide more qualified

and accurate judgments within their domain of expertise (Burgman et al. 2011). In some domains, such as weather forecasting, self-proclaimed experts are highly accurate (Wallsten and Budescu 1983). However, self-identified or peer-assessed expertise may not always be a reliable predictor of performance (Tetlock 2005; Burgman et al. 2011). Expertise isn't always easy to identify—it can be defined in a number of different ways, including experience, qualifications, performance on knowledge tests, and behavioral characteristics (Shanteau et al. 2002). Procedures to identify experts can lead to mathematical combination approaches that favor better, wiser, more expert judgments when judgments from multiple experts are available (French 1985, 2011; Budescu and Rantilla 2000; Aspinall 2010; Wang et al. 2011). In the subsections that follow, we discuss a number of general approaches that have been developed to assess the relative expertise in weighted averages and model-based aggregation procedures.

Performance weighting

A classic approach to aggregate expert opinions is based on Cooke's method (Cooke 1991; Bedford and Cooke 2001; Aspinall 2010). Cooke's method requires an independent stand-alone set of seed questions (sometimes referred to as calibration or control questions) with answers known to the aggregator but unknown to the experts. On the basis of performance on these seed questions, weights are derived that can be used to up-weight or down-weight experts' opinions on the remaining questions that don't have known answers (at least at the time of the experiment). Aspinall (2010) gives several real-world examples of Cooke's method, such as estimating failure times for dams exposed to leaks. Previous evaluations of Cooke's method may have led to over-optimistic results because the same set of seed questions used to calculate the performance weights were also used to evaluate model performance (Lin and Cheng 2009). Using a cross-validation procedure, Lin and Cheng (ibid.) showed that the performance-weighted average and an unweighted linear opinion pool in which all experts were equally weighted performed about the same. They concluded that it is wasn't clear whether the cost of generating and evaluating seed questions was justifiable. Liu et al. (2013) performed a theoretical analysis in a scenario in which the total number of questions that could be asked of judges was limited (e.g., each judge could estimate only fifty quantities), so that any introduction of seed questions necessarily reduced the number of questions with unknown ground truth (the questions of

ultimate interest). They found that under some conditions a small number of seed questions sufficed to evaluate the relative expertise of judges and measure any systematic response biases.

Using performance weighting, Budescu and Chen (2014) developed a contribution-weighted model in which the goal was to weight individuals by their contribution to the crowd in terms of the difference of the predictive accuracy of the crowd's aggregate estimate with, and without the judge's estimate in a series of forecasting questions. Therefore, individuals with a high contribution were those for which group performance would suffer if their judgment were omitted from the group average.

Generally, performance-based methods have the disadvantage that it can take time to construct seed questions with a known answer. As Shanteau et al. (2002) argued, experts may be needed in exactly those situations in which correct answers are not readily available. In forecasting situations, obvious choices for seed questions include forecasting questions that resolve during the time period over which the judge is evaluated. However, such procedures require an extended time commitment from judges and thus may not be practical in some scenarios.

Subjective Confidence

Another approach is to weight judgments by the subjective confidence expressed by the judges. In many domains, subjective confidence often demonstrates relatively low correlation with performance and accuracy (see, e.g., Tversky and Kahneman 1974; Mabe and West 1982; Stankov and Crawford 1997; Lee et al. 2012). However, in some cases a judge's confidence can be a valid predictor of accuracy. For example, in a group consisting of two people, a simple strategy of selecting the judgment of the more confident person (Koriat 2012) leads to better performance than relying on any single judgment. Koriat argues that subjective confidence may be driven more by common knowledge than by the correctness of the answer. However, it is possible to set up tasks in which the popular answer, typically associated with high confidence, is also the incorrect answer (Prelec and Seung 2006). Overall, performance from confidence-weighted judgments will depend heavily on the nature of the task and the degree to which the task is a representative sample of individuals (Hertwig 2012).

Coherence and Consistency

Coherence in probability judgments can be taken as a plausible measure of a judge's competence in probability and logic. Wang et al. (2011) and

Olson and Karvetski (2013) showed that down-weighting judgments of individuals associated with less coherent judgments (across questions) was effective in forecasting election outcomes. A related idea is that experts should produce judgments that are consistent over time such that similar responses are given to similar stimuli (Einhorn 1972, 1974). The within-person reliability or consistency can be used as a proxy for expertise, especially when it is combined with other cues for expertise such as discrimination (Shanteau et al. 2002; Weiss and Shanteau 2003; Weiss et al. 2009). One potential problem is that consistency is often assessed over short time intervals and with stimuli that are relatively easy to remember. In these cases, memory-retrieval strategies may limit the usefulness of consistency measures. Miller and Steyvers (2014) studied cases involving judgments that were difficult to remember explicitly and showed that consistency across repeated problems was strongly correlated with accuracy and that a consistency-weighted average of judgments was an effective aggregation strategy that outperformed the unweighted average.

Consensus-Based Models

The idea behind consensus-based models is that in many tasks the central tendency of a group leads to accurate answers. This group answer can be used as an estimate of the true answer to score individual members of a group. Individuals who produce judgments that are closer to the group's central tendency (across several questions) can be assumed to be more knowledgeable. Consensus-based models can therefore be used to estimate the knowledge of individuals in the absence of a known ground truth.

Consensus measures have been used in weighted averages where the judgments from consensus-agreeing individuals are up-weighted (Shanteau et al. 2002; Wang et al. 2011). Comprehensive probabilistic models for consensus-based aggregation were developed in the context of cultural consensus theory (Romney et al. 1987; Batchelder and Romney 1988) as well as in the context of observer-error models (Dawid and Skene 1979). To understand the basic approach, consider a scenario in which an observer has to figure out how to grade a multiple-choice test for which the answer key is missing. A consensus model posits a generative process in which each test taker, for each question, gives an answer that is a sample taken from a distribution in which the mean is centered on the latent answer key and the variance is treated as a variable that relates inversely to the latent ability of the observer.

Probabilistic inference can be used to simultaneously infer the answer key and the ability of each individual. Test takers with high ability are closer to the answer key on average; test takers with lower ability tend to deviate more from the answer key and from their higher-ability compatriots.

This consensus-based approach is not limited to problems for which the responses are discrete; it can also be used to estimate group responses over a continuous range of potential answers (Batchelder and Anders 2012). Consensus-based models are also able to account for variations in the difficulty of the questions. Consensus-based methods have led to many statistical models for crowdsourcing applications in which workers provide subjective labels for simple stimuli such as images (see, e.g., Smyth et al. 1995; Karger et al. 2011). Hierarchical Bayesian extensions have been proposed by Lipscomb et al. (1998) and by Albert et al. (2012).

Recently, consensus-based aggregation models have been applied to more complex decision tasks, such as ranking data (Lee et al. 2012, 2014). For example, individuals ranked a number of U.S. presidents in chronological order, or cities by their number of inhabitants. A simple generative model was proposed in which the observed ranking was based on the ordering of samples from distributions centered on the true answer but with variances determined by latent expertise levels. Lee et al. (2012) showed that the expertise levels inferred by the model were better correlated with actual performance than subjective confidence ratings provided by the participants.

Generally, consensus-based methods perform well on tasks that individuals do reasonably well (Weiss et al. 2009). One potential weakness of consensus-based methods is that they are vulnerable to cases in which agreement arises for reasons other than expertise. This can occur in challenging tasks on which the majority of individuals use heuristics. For example, when predicting the outcomes of certain sports tournaments, individuals who do not closely follow those tournaments may adopt heuristic strategies based on the familiarity of the teams (see, e.g., Goldstein and Gigerenzer 2002). Another potential issue is that in some cases it may be inappropriate to assume that there is a single latent answer or opinion for the whole group—there may be multiple clusters of individuals with divergent beliefs. In this case, consensus-based models must be extended to make inference over multiple groups with multiple answer keys; there has been preliminary work that shows that this may indeed be feasible (Anders and Batchelder 2012).

The Role of Meta-Cognition

The Bayesian Truth Serum (BTS) proposed by Prelec (2004) is an idea that incorporates metaknowledge—the knowledge of other people's judgments in aggregation. The BTS method was designed as an incentive mechanism to encourage truthful reporting. It can elicit honest probabilistic judgments even in situations in which the objective truth is difficult to obtain or intrinsically unknowable. It has been used to encourage people to answer survey questions truthfully (Weaver and Prelec 2013) and to estimate the prevalence of questionable research practices (John et al. 2012). However, it has also been tested in preliminary experiments on general knowledge questions (Prelec and Seung 2006) with which the performance of the method can be assessed objectively—for example, whether Chicago is the capital of Illinois. A minority of respondents might be expected to know the correct answer to that question. The majority of respondents might use simple heuristics that would lead them to the plausible yet incorrect answer. In the BTS approach, judges provide a private answer to a binary question and an estimate of the percentage of people who would give each response. The latter estimate involves metacognitive knowledge of other people. For each judge, a BTS score is calculated that combines the accuracy of the metacognitive judgments (rewarding an accurate prediction of other people's responses) and an information score that rewards surprisingly common responses. In the previous question, the correct answer—Springfield—will receive a high score if more people actually produced that answer than was predicted (metacognitively). Prelec and Seung (2006) showed that the BTS-weighted aggregate outperformed majority voting in a number of cases—cases in which only a minority of judges knew the correct answer. Though these initial empirical results are promising, it isn't clear how the BTS method will perform in areas, such as forecasting, in which the true answer isn't knowable at the time the question is asked, and metacognition about other people's forecasts might be biased in a number of ways. Recent research has also suggested that the metacognitive efficacy of individuals is positively correlated with the group's overall collective intelligence ability (Engel et al. 2014).

The Role of Information Sharing

Until recently, much of the work that has been done in collective decision making has involved a good deal of dynamic interaction among

group members (see, e.g., Lorge et al. 1958). Often a group of properly trained people with a lot of experience in working together can make judgments that are more accurate than those of any of the individual members (Watson et al. 1991; also see the chapters in this volume on Organizational Behavior and on Law, Communications, Sociology, Political Science, and Anthropology). When members of a group haven't been specifically trained to work together, the results can be far more varied; group members may have difficulty coordinating their responses to obtain a consensus (Steiner 1972; Lorenz et al. 2011) and are more vulnerable to cognitive biases and errors (Janis 1972; Stasser and Titus 1987; Kerr et al. 1996). It has been suggested that groups in which the individuals interact are most effective when their collective decision is arrived at by a weighted average of each member's opinions (Libby et al. 1987).

One popular method for soliciting group judgments is the Delphi method (Rowe and Wright 1999). By separating individuals, having them individually answer guided questionnaires, and allowing them to view one another's responses and to provide updated feedback, the Delphi method allows individuals to weight their own expertise in relationship to others and (ideally) to provide better-informed estimates. These individual estimates are then combined via statistical aggregation similar to the previous methods discussed. As with the training of specialized decision-making groups, there is still a large cost associated with setting up and coordinating a Delphi-based decision process. There are a number of additional schemes for limited information sharing that avoid many of the social and cognitive biasing that is inherent in dynamic group decision making (Gallupe et al. 1991; Olson et al. 2001; Whitworth et al. 2001; Rains 2005—also see the chapters in this volume on Human–Computer Interaction and Artificial Intelligence).

The effect of information sharing is strongly dependent on the type of network structure in which participants share information with one another (Kearns et al. 2006, 2012; Mason et al. 2008; Judd et al. 2010; Bernstein et al. 2011). For example, Mason et al. (2008) studied problem-solving tasks in which participants (corresponding to nodes on a network) were arranged in a number of different networks, some of them fully connected, some of them lattices, some of them random, and some of them small-world networks. The task for participants was to find the maximum of a continuous function with one input variable. Participants could probe the function with numerical values for the

input variable and obtain feedback by the value returned by the function. The function was sufficiently complicated with multiple local modes such that no individual could cover the space of possibilities within a reasonable amount of time. Participants received information about their neighbors' guesses and outcomes. The results showed that the network configuration had a strong effect on overall performance. Individuals found good global solutions more quickly in the small-world networks, relative to lattices and random networks, presumably because information can spread very quickly in small-world networks. It is not entirely clear why the small-world networks performed better than the fully connected networks, however. In a fully connected network, participants have full information about all other participants, and theoretically they should be able to benefit from that information. Mason et al. (2008) proposed that "less is more" in small-world networks—that participants may be better able to pay attention to the information from a smaller number of neighbors.

Kearns et al. (2006) and Judd et al. (2010) studied decentralized coordination games on networks in which each participant solved only a small part of a global problem. In contrast with the study by Mason et al. (2008), individuals were required to coordinate their efforts in order to collectively produce a good global solution. One coordination game involved a coloring problem in which each participant was required to choose a color from a fixed set of colors that was different from those of his or her neighbors. The results showed that the network structure had a strong influence on solution times. Long-distance connections hurt performance on the coloring task. On the other hand, if the task was altered so that consensus solutions were rewarded (i.e., all nodes had the same color), the long-distance connections improved performance. Across many of these coordination tasks, performance of human subjects came close to the optimal solution. Kearns et al. (2012) reported that 88 percent of the potential rewards available to human subjects were collected.

Task sharing can also be beneficial when individuals must explore a large problem space to find good solutions. Khatib et al. (2011) used a collective problem-solving approach to scientific discovery to optimize protein folding. Each player manipulated the protein folding to find stable configurations. One group of participants found a breakthrough solution to the problem that then was adopted by other participants as a new starting point for their own solutions. Collaboration also makes sense when questions are sufficiently complex that subjects

may have different parts of the answer (Malone et al. 2010). Miller and Steyvers (2011) explored rank-ordering tasks; the first subject in the task was given a random list ordering, and then each subject received the final ordering of a previous participant in an iterative fashion. In contrast with simpler information-passing tasks (see Beppu and Griffiths 2009), answers didn't necessarily converge on the correct ordering, but by aggregating across all subjects in the sequence it was possible to combine the partial knowledge of individuals into a nearly complete whole. Tools exist whereby this information sharing can be utilized to explicitly create external, shared collective knowledge as an aid to the collaborative process (Ren and Argote 2011). It has been shown that subjects are more susceptible to memory bias when given the responses of other subjects, but this can be overcome by using aggregation (Ditta and Steyvers 2013).

Collective Intelligence within Individuals

Whereas collective intelligence is often considered at the level of groups, we can also consider collective intelligence within an individual. In one experiment, Vul and Pashler (2008) asked individuals to estimate quantities (e.g., "What percentage of the world's airports are in the United States?") multiple times at varying time intervals. They found a "wisdom of the crowd" effect within one mind—the average of two guesses (from the same person) was more accurate than either of the individual guesses. This effect was larger if more time elapsed between the two estimates, presumably because participants' answers were less correlated because of strategic or memory effects. Hourihan and Benjamin (2010) found that the average of two guesses from individuals with short working-memory spans was more accurate than the average of two guesses from individuals with long working-memory spans, which suggested that the ability to remember the first response (as opposed to reconstructing an answer from general knowledge) might be an impediment to the "wisdom within one mind" effect.

Rauhut and Lorenz (2011) generalized Vul and Pashler's finding and demonstrated that the average over five repeated estimates was significantly better than the average from two repeated estimates (or a single estimate). This is somewhat surprising—one might assume that the first guess would already be based on all available information and that the subsequent guesses would not provide additional information. These findings show that there is an independent error component in

the estimates that can be canceled by averaging. Generally, these findings also support the concept that subjective estimates arise as samples from probabilistic representations underlying perceptual and cognitive models (Gigerenzer et al. 1991; Fiser et al. 2010; Griffiths et al. 2012).

The exact procedure used to elicit repeated judgments has been found to influence the "wisdom within one mind" effect. For example, a method known as dialectical bootstrapping (Herzog and Hertwig 2009) is designed to facilitate the retrieval of independent information from memory. Participants are told that their first estimate is off the mark and are asked to consider knowledge that was previously overlooked, ignored, or deemed inconsistent with current beliefs. Herzog and Hertwig showed that dialectical bootstrapping led to higher accuracy than standard instructions. In the More-Or-Less-Elicitation (MOLE) method (Welsh et al. 2009), participants are asked to make repeated relative judgments in which they have to select which of two options they think is closer to the true value. The advantage of this procedure is that it avoids asking the exact same question, which might elicit an identical answer.

Discussion

Human cognition plays a major role in the formation of collective intelligence by groups. In order to understand the collective intelligence of groups, we need to understand how judgments made by individual minds are affected by errors, biases, strategies, and task considerations. By developing aggregation methods and models that correct for these factors, and by using debiasing procedures in which individuals are trained to avoid such mistakes, it is possible to make more intelligent collective decisions.

It is also necessary to understand how the collective performance of a group is affected by the group's composition, by the relative expertise of the members, and by the sharing of information (if there is any such sharing) by the members. Such an understanding can help us to identify individuals who tend to produce more accurate judgments and also can help us to determine how and when to allow individuals to share information so as to make better collective estimates. In addition, by understanding the meta-cognition of the individuals in a group—their understanding of the other individuals—we can learn more about an individual's knowledge than we can learn from that individual's judgments alone.

Finally, one of the most important roles for cognitive research is to further our understanding of individuals' mental representations that are used to produce judgments. Converging evidence suggests that human knowledge is inherently probabilistic. Not only does this affect how individuals retrieve information from themselves; it also affects how they view others' information. The nature of these mental representations has implications for the kinds of aggregation models that are effective in combining human judgments, and for how collective intelligence arises generally.

References

Albert, I., S. Donnet, C. Guihenneuc-Jouyaux, S. Low-Choy, K. Mengersen, and J. Rousseau. J. 2012. Combining expert opinions in prior elicitation. *Bayesian Analysis* 7 (3): 503–532.

Anders, R., and W. H. Batchelder. 2012. Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology* 56 (6): 452–469.

Ariely, D., et al. 2000. The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology. Applied* 6 (2): 130.

Aspinall, W. 2010. A route to more tractable expert advice. *Nature* 463: 264–265.

Batchelder, W. H., and R. Anders. 2012. Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology* 56 (5): 316–332.

Batchelder, W. H., and A. K. Romney. 1988. Test theory without an answer key. *Psychometrika* 53 (1): 71–92.

Bedford, T., and R. Cooke. R. 2001. *Probabilistic Risk Analysis: Foundations and Methods*. Cambridge University Press.

Beppu, A., and T. L. Griffiths. 2009. Iterated learning and the cultural ratchet. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Bernstein, M. S., M. S. Ackerman, E. H. Chi, and R. C. Miller. 2011. The trouble with social computing systems research. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM.

Budescu, D., and E. Chen. 2014. Identifying expertise to extract the wisdom of crowds. *Management Science* 61 (2): 267–280.

Budescu, D. V., and A. K. Rantilla. 2000. Confidence in aggregation of expert opinions. *Acta Psychologica* 104 (3): 371–398.

Burgman, M. A., M. McBride, R. Ashton, A. Speirs-Bridge, L. Flander, B. Wintle, F. Fidler, L. Rumpff, and C. Twardy. 2011. Expert status and performance. *PLoS ONE* 6 (7): e22998.

Christensen-Szalanski, J., and J. B. Bushyhead. 1981. Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance* 7 (4): 928–935.

Cooke, R. M. 1991. *Experts in Uncertainty*. Oxford University Press.

Davis-Stober, C., D. Budescu, J. Dana, J., and S. Broomell. 2014. When is a crowd wise? *Decision* 1 (2): 79–101.

Dawid, A. P., and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics* 28 (1): 20–28.

Ditta, A. S., and M. Steyvers. 2013. Collaborative memory in a serial combination procedure. *Memory* 21 (6): 668–674.

Einhorn, H. J. 1972. Expert measurement and mechanical combination. *Organizational Behavior and Human Performance* 7: 86–106.

Einhorn, H. J. 1974. Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology* 59: 562–571.

Engel, D., A. W. Woolley, L. X. Jing, C. F. Chabris, and T. W. Malone. 2014. Reading the mind in the eyes or reading between the lines? Theory of mind predicts collective intelligence equally well online and face-to-face. *PLoS ONE* 9 (12): e115212.

Fiser, J., P. Berkes, G. Orbán, and M. Lengyel. 2010. Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Sciences* 14 (3): 119–130.

French, S. 1985. Group consensus probability distributions: A critical survey. In *Bayesian Statistics*, volume 2, ed. J. Bernardo, M. DeGroot, D. Lindley, and A. Smith. North-Holland.

French, S. 2011. Expert judgement, meta-analysis and participatory risk analysis. *Decision Analysis* 9 (2): 119–127.

Gallupe, R. B., L. M. Bastianutti, and W. H. Cooper. 1991. Unblocking brainstorms. *Journal of Applied Psychology* 76 (1): 137–142.

Gilovich, T. D. Griffin, and D. Kahneman, eds. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

Gigerenzer, G., U. Hoffrage, and H. Kleinbölting. 1991. Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review* 98: 506–528.

Goldstein, D. G., and G. Gigerenzer. 2002. Models of ecological rationality: The recognition heuristic. *Psychological Review* 109 (1): 75–90.

Griffiths, T. L., E. Vul, and A. N. Sanborn. 2012. Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science* 21 (4): 263–268.

Hertwig, R. 2012. Tapping into the wisdom of the crowd—with confidence. *Science* 336: 303–304.

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

Hogarth, R. M. 1975. Cognitive processes and the assessment of subjective probability distributions. *Journal of the American Statistical Association* 70 (35): 271–289.

Hourihan, K. L., and A. S. Benjamin. 2010. Smaller is better (when sampling from the crowd within): Low memory span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 36: 1068–1074.

Janis, I. L. 1972. *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin.

John, L. K., G. Loewenstein, and D. Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23 (5): 524–532.

Judd, S., M. Kearns, and Y. Vorobeychik. 2010. Behavioral dynamics and influence in networked coloring and consensus. *Proceedings of the National Academy of Sciences* 107 (34): 14978–14982.

Kahneman, D., P. Slovic, and A. Tversky. 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Kahneman, D., and A. Tversky, eds. 2000. *Choices, Values, and Frames*. Cambridge University Press.

Karger, D. R., S. Oh, and D. Shah. 2011. Iterative learning for reliable crowdsourcing systems. *Advances in Neural Information Processing Systems* 24: 1953–1961.

Kearns, M., S. Suri, and N. Montfort. 2006. An experimental study of the coloring problem on human subject networks. *Science* 313 (5788): 824–827.

Kearns, M., S. Judd, and Y. Vorobeychik. 2012. Behavioral experiments on a network formation game. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM.

Kerr, N. L., R. J. MacCoun, and G. P. Kramer. 1996. Bias in judgment: Comparing individuals and groups. *Psychological Review* 103 (4): 687–719.

Khatib, F., S. Cooper, M. D. Tyka, K. Xu, I. Makedon, Z. Popović, and F. Players. 2011. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences* 108 (47): 18949–18953.

Koriat, A. 2012. When are two heads better than one and why? *Science* 336 (6079): 360–362.

Lee, M. D., E. Grothe, and M. Steyvers. 2009. Conjunction and disjunction fallacies in prediction markets. In *Proceedings of the 31th Annual Conference of the Cognitive Science Society*, ed. N. Taatgen, H. van Rijn, L. Schomaker, and J. Nerbonne. Erlbaum.

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

Lee, M. D., M. Steyvers, and B. Miller. 2014. A cognitive model for aggregating people's rankings. *PLoS ONE* 9 (5): e96431.

Lee, M. D., S. Zhang, and J. Shi. 2011. The wisdom of the crowd playing *The Price Is Right*. *Memory & Cognition* 39 (5): 914–923.

Libby, R., K. T. Trotman, and I. Zimmer. 1987. Member variation, recognition of expertise, and group performance. *Journal of Applied Psychology* 72 (1): 81–87.

Lichtendahl, K. C., Y. Grushka-Cockayne, and P. Pfeifer. 2013. The wisdom of competitive crowds. *Operations Research* 61 (6): 1383–1398.

Lin, S. W., and C. H. Cheng. 2009. The reliability of aggregated probability judgments obtained through Cooke's classical model. *Journal of Modelling in Management* 4 (2): 149–161.

Lipscomb, J., G. Parmigiani, and V. Hasselblad. 1998. Combining expert judgment by hierarchical modeling: An application to physician staffing. *Management Science* 44: 149–161.

Liu, Q., M. Steyvers, and A. Ihler. 2013. Scoring workers in crowdsourcing: How many control questions are enough? *Advances in Neural Information Processing Systems* 26: 1914–1922.

Lorenz, J., H. Rauhut, F. Schweitzer, and D. Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences* 108 (22): 9020–9025.

Lorge, I., D. Fox, J. Davitz, and M. Brenner. 1958. A survey of studies contrasting the quality of group performance and individual performance, 1920–1957. *Psychological Bulletin* 55 (6): 337–372.

Mabe, P. A., and S. G. West. 1982. Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology* 67 (3): 280–296.

Malone, T. W., R. Laubacher, and C. Dellarocas. 2010. The Collective Intelligence Genome. *Sloan Management Review* 51 (3): 21–31.

Mandel, D. R. 2005. Are risk assessments of a terrorist attack coherent? *Journal of Experimental Psychology: Applied* 11 (4): 277–288.

Mason, W. A., A. Jones, and R. L. Goldstone. 2008. Propagation of innovations in networked groups. *Journal of Experimental Psychology. General* 137 (3): 422–433.

Massey, C., J. P. Simmons, and D. A. Armor. 2011. Hope over experience: Desirability and the persistence of optimism. *Psychological Science* 22 (2): 274–281.

Mellers, B., L. Ungar, J. Baron, J. Ramos, B. Gurcay, K. Fincher, S. E. Scott, et al. 2014. Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science* 25 (5): 1106–1115.

Miller, B. J., and M. Steyvers. 2011. The wisdom of crowds with communication. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Miller, B. J., and M. Steyvers. 2014. Improving Group Accuracy Using Consistency across Repeated Judgments. Technical report, University of California, Irvine.

Norman, D. A. 1993. Distributed cognition. In *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*, ed. T. Dunaeff and D. Norman. Perseus Books.

Olson, K. C., and C. W. Karvetski. 2013. Improving expert judgment by coherence weighting. In *Proceedings of 2013 IEEE International Conference on Intelligence and Security Informatics*. IEEE.

Olson, G. M., T. W. Malone, and J. B. Smith, eds. 2001. *Coordination Theory and Collaboration Technology*. Erlbaum.

Ottaviani, M., and P. N. Sørensen. 2006. The strategy of professional forecasting. *Journal of Financial Economics* 81 (2): 441–466.

Prelec, D. 2004. A Bayesian truth serum for subjective data. *Science* 306: 462–466.

Prelec, D., and H. S. Seung. 2006. An algorithm that finds truth even if most people are wrong. Unpublished manuscript.

Rains, S. A. 2005. Leveling the organizational playing field—virtually: A meta-analysis of experimental research assessing the impact of group support system use on member influence behaviors. *Communication Research* 32 (2): 193–234.

Rauhut, H., and J. Lorenz. 2011. The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology* 55 (2): 191–197.

Ren, Y., and L. Argote. 2011. Transactive memory systems 1985–2010: An integrative framework of key dimensions, antecedents, and consequences. *Academy of Management Annals* 5 (1): 189–229.

Romney, A. K., W. H. Batchelder, and S. C. Weller. 1987. Recent applications of cultural consensus theory. *American Behavioral Scientist* 31 (2): 163–177.

Rowe, G., and G. Wright. 1999. The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting* 15 (4): 353–375.

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

Shanteau, J., D. J. Weiss, R. P. Thomas, and J. C. Pounds. 2002. Performance-based assessment of expertise: How to decide if someone is an expert or not. *European Journal of Operational Research* 136 (2): 253–263.

Simmons, J. P., L. D. Nelson, J. Galak, and S. Frederick. 2011. Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research* 38 (1): 1–15.

Smyth, P., U. Fayyad, M. Burl, P. Perona, and P. Baldi. 1995. Inferring ground truth from subjective labeling of Venus images. *Advances in Neural Information Processing Systems*: 1085–1092.

Stankov, L., and J. D. Crawford. 1997. Self-confidence and performance on tests of cognitive abilities. *Intelligence* 25 (2): 93–109.

Stasser, G., and W. Titus. 1987. Effects of information load and percentage of shared information on the dissemination of unshared information during group discussion. *Journal of Personality and Social Psychology* 53 (1): 81–93.

Steiner, I. D. 1972. *Group Process and Productivity*. Academic Press.

Steyvers, M., M. D. Lee, B. Miller, and P. Hemmer. 2009. The wisdom of crowds in the recollection of order information. *Advances in Neural Information Processing Systems* 22: 1785–1793.

Steyvers, M., T. S. Wallsten, E. C. Merkle, and B. M. Turner. 2014. Evaluating probabilistic forecasts with Bayesian signal detection models. *Risk Analysis* 34 (3): 435–452.

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

Taleb, N. N. 2007. *The Black Swan: The Impact of the Highly Improbable*. Random House.

Tetlock, P. E. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press.

Turner, B. M., M. Steyvers, E. C. Merkle, D. V. Budescu, and T. S. Wallsten. 2014. Forecast aggregation via recalibration. *Machine Learning* 95 (3): 261–289.

Tversky, A., and D. Kahneman. 1974. Judgment and uncertainty: Heuristics and biases. *Science* 185: 1124–1131.

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

Wallsten, T. S., and D. V. Budescu. 1983. State of the art—Encoding subjective probabilities: A psychological and psychometric review. *Management Science* 29 (2): 151–173.

Wallsten, T. S., D. V. Budescu, L. Erev, and A. Diederich. 1997. Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making* 10: 243–268.

Wang, G., S. R. Kulkarni, H. V. Poor, and D. N. Osherson. 2011. Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis* 8 (2): 128–144.

Watson, W. E., L. K. Michaelsen, and W. Sharp. 1991. Member competence, group interaction, and group decision making: A longitudinal study. *Journal of Applied Psychology* 76 (6): 803–809.

Weaver, R., and D. Prelec. 2013. Creating truth-telling incentives with the Bayesian Truth Serum. *Journal of Marketing Research* 50 (3): 289–302.

Wegner, D. M. 1987. Transactive memory: A contemporary analysis of the group mind. In *Theories of Group Behavior*, ed. B. Mullen and G. Goethals Springer.

Weiss, D. J., and J. Shanteau. 2003. Empirical assessment of expertise. *Human Factors* 45 (1): 104–116.

Weiss, D. J., K. Brennan, R. Thomas, A. Kirlik, and S. M. Miller. 2009. Criteria for performance evaluation. *Judgment and Decision Making* 4 (2): 164–174.

Welsh, M. B., M. D. Lee, and S. H. Begg. 2009. Repeated judgments in elicitation tasks: Efficacy of the MOLE method. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Whitworth, B., B. Gallupe, and R. McQueen. 2001. Generating agreement in computer-mediated groups. *Small Group Research* 32 (5): 625–665.