COGNITION AND COLLECTIVE INTELLIGENCE

Mark Steyvers and Brent Miller

University of California, Irvine

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 and decision making, and problem solving – including meta-cognitive tasks, when an individual is reflecting on their own or other people's performance. Understanding these 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. Each of these components alone, or in concert, can be understood to form the basic building blocks of group collective intelligence.

Consider the classic estimation task where a group of individuals must determine the number of marbles in 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 a better answer than that of most of the individuals. This has come to be known as the wisdom of crowds effect (Ariely et al, 2000; Davis-Stober, Budescu, & Broomell, in press; Surowiecki, 2004; Wallsten, Budescu, Erev, & Diederich, 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. If we start to make more realistic assumptions about the estimation task or change it to make it more like complex, real-world situations however, it may quickly become obvious how psychological factors can come into play. Suppose for example that individuals give judgments that are systematically biased (e.g., people might overestimate the number of marbles due to the comparative difference in size). How can we know what the potential biases are, and how to correct for them? Suppose that a given individual is better at the task than others, doesn't understand the task, or isn'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 reasoning, how does this information sharing affect the results? To fully understand how collective intelligence arises from a group of individuals, and how to improve the group's collective wisdom, it is critical 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 the issue of expertise, and discuss 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 such as the Bayesian truth serum that identify knowledgeable individuals in the absence of any ground truth. We will then look at how information sharing between individuals affects the collective performance, and review a number of studies that

manipulate how that information is shared. Finally, we will look at collective intelligence within a single mind.

Identifying and Correcting for Biases

The literature has many examples of cognitive biases that can systematically distort individual human judgment (Hogarth, 1975; Kahneman, Slovic, Tversky, 1982). For example, human probabilistic judgments can be over-confident about reported probabilities, neglect the event's base rate or be biased by the desirability of the outcomes (Kahneman & Tversky, 2000; Gilovich, Griffin & Kahneman, 2002; Massey, Simmons, & Armor, 2011). Conversely, individuals may often be sensitive to extraneous information that can be irrelevant to the judgment task at hand (Goldstein & Gigerenzer, 2002). Systematic distortions that affect individual judgments can also affect group performance. Although uncorrelated errors at the level of individual judgments can be expected to average out in the group average, systematic biases and distortions cannot be averaged out by using standard statistical averaging approaches (Simmons et al., 2011; Steyvers, Wallsten, Merkle, Turner, 2014).

It has been shown that by training individuals in the potential biases of estimation, it is possible to get subjects to debias their own estimates, at least to a certain degree (Mellers et al., 2014). Alternatively, by understanding what these cognitive biases are, it is possible 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 & Bushyhead, 1981). In judging other events that occur more frequently, such as in weather forecasting, experts have more opportunity to properly calibrate their responses (Wallsten & 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. (2013) 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 a 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 upon each other, the result can lead to incoherent probability judgments that do not follow the rules of probability theory (Wang, Kulkarni, Poor, & Osherson, 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 might not always connect these interrelations in logical ways and might fail to produce coherent probability judgments. Failure of coherence can occur at the individual level (e.g. Mandel, 2005), but also at the aggregate level in prediction markets (Lee, Grothe, Steyvers, 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 might persist even in the presence of financial incentives (Lee, Grothe, & Steyvers, 2009). Wang et al. (2011) proposed a

weighted coherentization approach that combines credibility weighting with coherentization, such that the aggregate judgments are guaranteed to obey the rules of probability. For instance, they asked participants to forecast 2008 US Presidential election outcomes and included questions of elementary events, but also 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 their estimates when the demands of the task require it. 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 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 biased from individuals' true estimates. On the game show *The Price is Right*, contestants bid in sequential order on the price of an item, where the winner is the closest without going over. Contestants often give estimates that are quite far from the actual price (and presumably their own beliefs), in order to increase their odds of winning the competition against their peers. Aggregation approaches that model the strategic considerations of these competitive environments and attempt to aggregate over inferred beliefs outperform standard aggregation methods (Lee, Zhang & Shi, 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 might have to make riskier but more informative estimates for aggregation (Lichtendahl et al., 2013).

Identifying Expert Judgments

The ability to identify and use 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 (e.g, Wallsten & Budescu, 1983). However, self-identified or peer-assessed expertise might not always be a reliable cue for performance (Tetlock, 2005; Burgman et al. 2011). Expertise is not always easy to identify because it can be defined in a number of ways, including experience, qualifications, knowledge tests, and behavioral characteristics (Shanteau, Weiss, Thomas & Pounds, 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 & Rantilla, 2000; Aspinall, 2010; Wang, Kulkarni, Poor & Osherson, 2011). Below, 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 & Cooke, 2001; Aspinal, 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- or down-weight the opinions on the remaining questions that don't have known answers (at least at time of the experiment). Aspinal (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 might have led to overly optimistic results because the same set of seed questions used to calculate the performance weights were also used to evaluate model performance (Lin & Chen, 2009). Using a cross-validation procedure, Lin and Chen (2009) showed that the performance-weighted average and an unweighted linear opinion pool in which all experts are equally weighted performed about the same. They concluded that it is unclear whether the cost of generating and evaluating seed questions is justifiable. Recently, Qiang, Steyvers, and Ihler (in press) performed a theoretical analysis in a scenario where the total number of questions that can be asked of judges is limited (e.g., each judge can only estimate 50 quantities). Therefore, any introduction of seed questions necessarily cuts down on 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 are sufficient to evaluate the relative expertise of judges and measure any systematic response biases.

In a recent performance-weighting approach, Budescu and Chen (under review) developed the contribution weighted model. In this approach, the goal is 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 are those for which group performance will suffer if their judgment is omitted from the group average.

Generally, performance-based methods have a disadvantage in that it can take time to construct seed questions with a known answer. As Shanteau et al. (2002) argued, experts might be needed in exactly those situations where correct answers are not readily available. In forecasting situations, an obvious choice for seed questions is the use of forecasting questions that resolve during the time period that the judge is evaluated. However, such procedures require an extended time commitment from judges that might 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 (e.g. Tversky and Kahneman, 1974; Mabe and West, 1982; Stankov & Crawford, 1997; Lee, Steyvers, Young & Miller, 2012). However, a judge's confidence can in some cases be a valid predictor of accuracy. For example, in a group involving two people, a simple strategy of selecting the judgment of the person with the highest confidence (Koriat, 2012) leads to better performance than relying on any of the individual judgments. Koriat argues that subjective confidence might be driven more by common knowledge as opposed to the correctness of the answer. It is possible to set up tasks where the popular answer, typically associated with high confidence, is also the incorrect answer (Prelec & Seung, 2006). Overall, performance from confidence weighted judgments will strongly depend on the nature of the task and the degree to which the task is a representative sample (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 downweighting 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 combined with other cues for expertise such as discrimination (Shanteau, Weiss, Thomas, & Pounds, 2002; Weiss & Shanteau, 2003; Weiss, Brennan, Thomas, Kirlik, & Miller, 2009). One potential problem is that consistency is often assessed over short time intervals and using stimuli that are relatively easy to remember. In these cases, memory retrieval strategies might limit the usefulness of consistency measures. Miller and Steyvers (submitted) studied cases involving judgments that are difficult to remember explicitly. They 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 and this group answer can be used as a proxy of the true answer to score individual group members - 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 knowledgeability 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 upweighted (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, Batchelder, & Weller, 1987; Batchelder & Romney, 1988) as well as observer-error models (Dawid & Skene, 1979). To understand the basic approach, consider a scenario where 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 which is a sample taken from a distribution where 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 as well as the abilities of each individual. Test-takers with high ability are closer to the answer key on average, and test-takers with a lower ability tend to deviate more from the answer key, and their higher-ability compatriots.

This consensus-based approach is not limited to problems where responses are discrete – it can also be used to estimate group responses over a continuous range of potential answers (Batchelder & Anders, 2012). Consensus-based models are also able to account for variations in the difficulty and challenges of

the questions themselves. Consensus-based methods have led to many statistical models for crowd-sourcing applications, where workers provide subjective labels for simple stimuli such as images (e.g. Smyth, Fayyad, Burl, Perona, & Baldi, 1995; Karger, Oh, & Shah, 2011). Hierarchical Bayesian extensions have been proposed by Lipscomb, Parmigiani, and Hasselblad (1998) and Albert, Donnet, Guihenneuc-Jouyaux, Low-Choy, Mengersen, and Rousseau (2012).

Recently, consensus-based aggregation models have been applied to more complex decision tasks, such as ranking data (Lee, Steyvers, de Young & Miller, 2012; Lee, Steyvers, & Miller, in press). 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 where 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 in tasks where people do reasonably well (Weiss et al., 2009). One potential problem with consensus-based methods is that it is vulnerable to cases where agreement arises for reasons other than expertise. This can occur in challenging tasks where the majority of individuals adopt heuristics to produce an answer. For example, when predicting the outcome of sports tournaments, individuals who do not closely follow these tournaments might adopt heuristic strategies based on the familiarity of the teams (e.g. Goldstein & Gigerenzer, 2002). Another potential problem is that in some cases it might be inappropriate to assume a single latent answer or opinion for the whole group – there might be multiple clusters of individuals with divergent beliefs. In this case, consensus-based models need to be extended to infer multiple groups with multiple answer keys; there has been preliminary work that shows that this may indeed be feasible (Anders & Batchelder, 2012).

Role of Meta Cognition

The Bayesian Truth Serum (BTS; Prelec, 2004) is a recent idea that incorporates metaknowledge, the knowledge of other'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 where the objective truth is intrinsically unknowable or difficult to obtain. It has been used to encourage people to answer truthfully in survey research (Weaver & Prelec, 2013) and to estimate the prevalence of questionable research practices (John, Loewenstein, Prelec, 2012). However, the method has also been tested in preliminary experiments on general knowledge questions (Prelec & Seung, 2006) where the performance of the method can be assessed objectively. In one example question, participants were asked whether Chicago was the capital of Illinois. This is a question where a minority of respondents might be expected to give the correct answer. The majority of respondents might use simple heuristics that lead them to the plausible yet incorrect answer. In the BTS approach, judges provide a private answer to a binary question, as well as 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) as well as an information score that rewards surprisingly common responses. In the capital of Illinois 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) show that the BTS-weighted aggregate outperforms majority voting in a number of cases --- essentially cases where only a minority of judges know the correct answer. While these initial empirical results are promising, it is unclear how the method will perform in areas such as forecasting, where the true answer in unknowable at the time the question is asked, and metacognition about other people's forecasts might be biased in a number of ways.

The Role of Information Sharing

Much of the work that has been done in collective decision making has historically involved a good deal of dynamic group interaction (Lorge et al., 1958). A group of people that is properly trained, and has a good deal of experience working together, can often make judgments that are more accurate than that of any of the individual members (Watson et al., 1991). When groups are not specifically trained to work together, the results can be far more varied however; group members can have trouble 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 & Titus, 1987; Kerr, MacCoun & Kramer, 1996). It has been suggested that interacting groups 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 & Wright, 1999). By separating individuals, having them individually answer guided questionnaires, and allowing them to view each other's responses and provide updated feedback, the Delphi method allows individuals to weight their own expertise in relationship to others and provide an (ideally) more-informed estimate. These individual estimates are then combined via statistical aggregation similar to those of non-communicative groups. As with training 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 inherent dynamic group decision making (Gallupe et al., 1991; Olson, Malone & Smith, 2001; Whitworth et al., 2001; Rains, 2005).

The effect of information sharing strongly depends on the type of network structure in which participants share information with each other (Kearns, Suri, & Montfort, 2006; Mason, Jones, and Goldstone, 2008; Judd, Kearns, & Vorobeychik, 2010; Bernstein et al., 2011; Kearns, Judd, & Vorobeychik, 2012). For example, Mason et al. (2008) studied problem-solving tasks where participants (corresponding to nodes on a network) were arranged in a number of different networks, including fully connected networks, lattices, random networks and 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 neighbor's guesses and outcomes. The results showed that the network

configuration had a strong impact 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 these networks. It is not entirely clear why the performance in the small-world network was better than the fully connected network, however. In a fully connected network, participants have full information about all other participants, and they theoretically should be able to benefit from this information. Mason et al. (2008) proposed that 'less is more' in these networks – participants might 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 where each participant solves only a small part of a global problem. Unlike the Mason et al. (2008) study, individuals were required to coordinate their efforts in order to collectively produce a good global solution. One coordination game involved a coloring problem where each participant needed to choose a color from a fixed set of colors that is different from their neighbors. The results showed that the network structure had a strong influence on solution times. Long-distance connections hurt performance in the coloring task. On the other hand, if the task was altered such that consensus solutions were rewarded (i.e., all nodes have 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 (2012) reported that 88% 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. In Khatib et al. (2011), a collective problem-solving approach to scientific discovery was used to optimize protein structures. Each player manipulated the protein folding structure to find stable configurations. One group of participants found a breakthrough solution to the problem that was adopted by other participants as a new starting point in their own solutions. Collaboration also makes sense when questions are complex enough that subjects may have different parts of the answer (Malone et al., 2009). In Miller & Steyvers (2011), rank ordering tasks were explored; 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. Unlike in simpler information passing tasks (see Beppu & Grifiths, 2009), answers did not necessarily converge on the correct ordering, but by aggregating across all subjects in the sequence, it was possible to combine the partial knowledge of each individual into a nearly-complete whole. It has been shown in subsequent experiments that subjects are more susceptible to memory bias when given the responses of another, but by aggregation this can be overcome (Ditta & Steyvers, 2013).

Collective Intelligence within Individuals

Whereas collective intelligence is often considered at the level of groups of individuals, we can also consider collective intelligence within a single individual. In one such 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 due to strategic or memory effects. Hourihan and Benjamin (2010) found the average of two guesses from individuals with low working-memory spans was more accurate than individuals with high working-memory spans, suggesting 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.

Similarly, Rauhut and Lorenz (2011) generalized this finding and demonstrated that the average over 5 repeated estimates was significantly better than the average from 2 repeated estimates (or a single estimate). This is somewhat surprising because 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 cancelled by averaging. Generally, they also support the concept that subjective estimates arise as samples from probabilistic representations underlying perceptual and cognitive models (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Fiser, Berkes, Orban, & Lengyel, 2010; Griffiths, Vul, & Sanborn, 2012).

The exact procedure used to elicit repeated judgments has been found to influence the wisdom-within one mind effect. For example, methods such as dialectical bootstrapping (Herzog and Hertwig, 2009) are designed to facilitate the retrieval of independent information from memory. In this procedure, 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 the dialectical procedure led to higher accuracy than standard instructions. Another procedure is the More-Or-Less-Elicitation (MOLE) method (Welsh, Lee, & Begg, 2009). Participants are asked to make repeated relative judgments where they have to select which of two options is thought to be 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 key role in the formation of group collective intelligence. In order to understand these groups, we need to understand how judgments from individual minds are affected by errors, biases, strategies and task considerations. By the development of aggregation models that correct for these factors, as well as debiasing procedures where 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 factors such as group composition, relative expertise among members and the information being sharing amongst the group (if any). This can help to identify individuals who tend to produce more accurate judgments and also shows us how and when to allow individuals to share information to make better collective estimates. Additionally, by understanding the meta-cognition of individuals – their understanding of the other individuals in a group – we can learn more about an individual's knowledge than just their estimate itself.

Finally, one of the most important roles for cognitive research is to better understand the mental representations that are used to produce the judgments. Converging evidence suggests that human knowledge and judgment is inherently probabilistic in nature. This affects not only how individuals retrieve information from themselves, but shapes how they view the information of others. The nature of these representations has implications for the kinds of aggregation models that are effective in combining human judgments.

References

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

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

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., 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., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, 56(5), 316-332.

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

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

Beppu, A., & Griffiths, T. L. (2009). Iterated learning and the cultural ratchet. *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2089-2094). Austin, TX: Cognitive Science Society.

Bernstein, M. S., Ackerman, M. S., Chi, E. H., & Miller, R. C. (2011). The trouble with social computing systems research. *CHI'11 Extended Abstracts on Human Factors in Computing Systems* (pp. 389-398). New York: ACM.

Budescu, D., & Chen, E. (under review). Identifying expertise and using it to extract the Wisdom of Crowds. *Management Science*.

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

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

Christensen-Szalanski, J., & Bushyhead, J. B. (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: Oxford University Press.

Davis-Stober, C., Budescu, D.V., Dana, J., Broomell, S. (in press). When is a crowd wise? *Decision*.

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

Ditta, A. S., & Steyvers, M. (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.

Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (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. J. M. Bernardo, M. H. DeGroot , D. V. Lindley, A. F. M. Smith, eds. *Bayesian Statistics*, vol. 2, North-Holland, Amsterdam, 183-201.

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

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

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, Mass.: Cambridge University Press.

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

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: the Recognition Heuristic. *Psychological Review,* 109(1), 75-90.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (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., & Hertwig, R. (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., Benjamin, A. S. (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*. Oxford: Houghton Mifflin.

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

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

Kahneman, D., Slovic, P., Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, Mass.: Cambridge University Press.

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge, Mass.: Cambridge University Press.

Karger, D. R., Oh, S., & Shah, D. (2011). Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems,* (pp. 1953-1961). Cambridge, Mass.: MIT Press.

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

Kearns, M., Judd, S., & Vorobeychik, Y. (2012). Behavioral experiments on a network formation game. In *Proceedings of the 13th ACM Conference on Electronic Commerce* (pp. 690-704). New York: ACM.

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

Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popović, Z., & Players, F. (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., Grothe, E., & Steyvers, M. (2009). Conjunction and Disjunction Fallacies in Prediction Markets. In N. Taatgen, H. van Rijn, L. Schomaker and J. Nerbonne (Eds.) *Proceedings of the 31th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

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

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., Steyvers, M., & Miller, B.J. (in press). A cognitive model for aggregating people's rankings. *PLoS ONE*. Accepted 8-Apr-2014.

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

Lichtendahl, K. C., Grushka-Cockayne, Y. & Pfeifer, P. (2013). The Wisdom of the Competitive Crowds, working paper.

Lin, S.-W., 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., Parmigiani, G., and Hasselblad, V. (1998). Combining expert judgment by hierarchical modeling: an application to physician staffing. *Management Science*, 44, 149-161.

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

Lorge, I., Fox, D., Davitz, J., & Brenner, M. (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., West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology*, 67(3), 280-296.

Malone, T., Laubacher, R., & Dellarocas, C. (2009). *Harnessing crowds: Mapping the genome of collective intelligence*. Cambridge, Mass.: MIT Sloan School of Management.

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

Massey, C., Simmons, J. P., & Armor, D. A. (2011). Hope Over Experience Desirability and the Persistence of Optimism. *Psychological Science*, 22(2), 274-281.

Mellers. B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S.E., Moore, D., Atanasov, P., Swift, S., Murray, T., & Tetlock, P. (2014). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science* (in press).

Miller, B. J., & Steyvers, M., (2011). The Wisdom of Crowds with Communication. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp1292-1297). Austin, TX: Cognitive Science Society.

Miller, B. J., & Steyvers, M., (submitted). Improving Group Accuracy by Using Consistency Across Repeated Judgments.

Olson, K. C., & Karvetski, C. W. (2013). Improving expert judgment by coherence weighting. In *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on* (pp. 197-199). IEEE.

Olson, G. M., Malone, T. W., & Smith, J. B. (Eds.). (2001). *Coordination theory and collaboration technology*. Mahwah, NJ: Erlbaum.

Ottaviani, M., & Sørensen, P. N. (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., & Seung, H. S. (2006). An algorithm that finds truth even if most people are wrong. Unpublished manuscript.

Liu, Q., Steyvers, M., & Ihler, A. (in press). Scoring Workers in Crowdsourcing: How Many Control Questions are Enough? *Advances in neural information processing systems*.

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., & Lorenz, J. (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.

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

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: issues and analysis. *International journal of forecasting*, 15(4), 353-375.

Shanteau, J., Weiss, D. J., Thomas, R. P., & Pounds, J. C. (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.

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.

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

Smyth, P., Fayyad, U., Burl, M., Perona, P., & Baldi, P. (1995). Inferring ground truth from subjective labelling of venus images. *Advances in neural information processing systems*, 1085-1092.

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

Stasser, G., & Titus, W. (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*. New York: Academic Press.

Steyvers, M., Lee, M. D., Miller, B., & Hemmer, P. (2009). The Wisdom of Crowds in the Recollection of Order Information. In Y. Bengio and D. Schuurmans and J. Lafferty and C. K. I. Williams and A. Culotta (Eds.) *Advances in neural information processing systems, 22*, pp. 1785-1793. Cambridge, Mass.: MIT Press.

Steyvers, M., Wallsten, T. S., Merkle, E.C., and Turner, B. M. (2014). Evaluating Probabilistic Forecasts with Bayesian Signal Detection Models. *Risk Analysis*, 34(3), 435-452.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Random House.

Tetlock, P.E. (2005). *Expert Political Opinion, How Good is it? How Can we Know*? Princeton: Princeton University Press.

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

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

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

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

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

Wang, G., Kulkarni, S.R., Poor, H.V., Osherson, D.N. (2011). Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis*, 8(2) 28–144.

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

Weaver, R., Prelec, D. (2013). Creating Truth-Telling Incentives with the Bayesian Truth Serum. *Journal of Marketing Research*, 50(3), 289-302.

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(1), 104-116.

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

Welsh, M. B., Lee, M. D., & Begg, S. H. (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: Austin, TX.

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