The Seventh AAAI Conference on Human
Computation and Crowdsourcing (HCOMP-19)

Studying the "Wisdom of Crowds" at Scale

Camelia Simoiu,[1] **Chiraag Sumanth,**[1] **Alok Mysore,**[2] **Sharad Goel**[1]
[1]Stanford University, [2]University of California San Diego

Abstract

In a variety of problem domains, it has been observed that the aggregate opinions of groups are often more accurate than those of the constituent individuals, a phenomenon that has been dubbed the "wisdom of the crowd". However, due to the varying contexts, sample sizes, methodologies, and scope of previous studies, it has been difficult to gauge the extent to which conclusions generalize. To investigate this question, we carried out a large online experiment to systematically evaluate crowd performance on 1,000 questions across 50 topical domains. We further tested the effect of different types of social influence on crowd performance. For example, in one condition, participants could see the cumulative crowd answer before providing their own. In total, we collected more than 500,000 responses from nearly 2,000 participants. We have three main results. First, averaged across all questions, we find that the crowd indeed performs better than the average individual in the crowd—but we also find substantial heterogeneity in performance across questions. Second, we find that crowd performance is generally more consistent than that of individuals; as a result, the crowd does considerably better than individuals when performance is computed on a full *set* of questions within a domain. Finally, we find that social influence can, in some instances, lead to herding, decreasing crowd performance. Our findings illustrate some of the subtleties of the wisdom-of-crowds phenomenon, and provide insights for the design of social recommendation platforms.

Introduction

Are crowds mad or wise? In his 1841 book, "Memoirs of extraordinary popular delusions and the madness of crowds," Charles Mackay documents a series of remarkable tales of human folly, ranging from the hysteria of the South Sea Bubble that ruined many British investors in the 1720s, to Holland's seventeenth-century "tulipomania", when individuals went into debt collecting tulip bulbs until a sudden depreciation in the bulbs' value rendered them worthless (Mackay 1841). Decades later, in yet another classic example, the statistician Francis Galton watched as eight hundred people competed to guess the weight of an ox at a county fair. He famously observed that the median of the guesses—1,207 pounds—was, remarkably, within 1% of the true weight (Galton 1907).

Over the past century, there have been dozens of studies that document this "wisdom of crowds" effect (Surowiecki 2005). Simple aggregation—as in the case of Galton's ox competition—has been successfully applied to aid prediction, inference, and decision making in a diverse range of contexts. For example, crowd judgments have been used to successfully answer general knowledge questions (Surowiecki 2005), identify phishing websites and web spam (Moore and Clayton 2008; Liu et al. 2012), forecast current political and economic events (Budescu and Chen 2014; Griffiths and Tenenbaum 2006; Hill and Ready-Campbell 2011), predict sports outcomes (Herzog and Hertwig 2011; Goel et al. 2010), and predict climate-related, social, and technological events (Hueffer et al. 2013; Kaplan, Skogstad, and Girshick 1950). However, given the diversity of experimental designs, subject pools, and analytic methods employed, it has been difficult to know whether these documented examples are a representative collection of a much larger space of tasks that exhibit a wisdom-of-crowds phenomenon, or conversely, whether they are highly specific instances of an interesting, though ultimately limited occurrence.

Moreover, it is unclear whether these findings generalize to many real-world settings where individuals make decisions under the influence of others' judgments. This question is especially relevant today, as peer influence is oftentimes explicitly built into online platforms. One might choose a restaurant, watch a movie, read a news story, or purchase a book because of the aggregated opinions of the "crowd." Recommender systems may display top-rated products first by default, whose quality has been estimated as the most popular or highly voted. In recent years, researchers have debated whether social influence undermines or enhances the wisdom of crowds. On the one hand, some have conjectured that if participants receive information about the answers of others, that can help ground responses, leading to greater accuracy (Faria et al. 2010; King et al. 2012; Madirolas and de Polavieja 2015). But, on the other hand, there is also worry that such social influence could result

in herding, which in turn could decrease collective performance (Lorenz et al. 2011; Muchnik, Aral, and Taylor 2013; Salganik, Dodds, and Watts 2006).

To systematically explore the wisdom-of-crowds phenomenon—including the effects of social influence—we carried out a large-scale, online experiment. In one of the most comprehensive studies of the wisdom-of-crowds effect to date, we collected a total of more than 500,000 responses to 1,000 questions across 50 topical areas. For each question, we computed the "crowd" answer by either taking the median response of participants (in the case of open-ended, numerical questions) or the most popular choice (in the case of categorical questions).

Averaged across our full set of questions, we found that the crowd answer was approximately in the 65th percentile of individual responses, ranked by accuracy. Our results thus lend support to the idea that the wisdom-of-crowds effect indeed holds on a corpus chosen to reflect a wide variety of topical areas. Further, we found that crowd performance was typically more consistent than the performance of individuals. That is, whereas the crowd performed at least modestly better than average on all of the questions, even the best individuals occasionally performed poorly. As a result, when we looked at performance at the level of topical domains, rather than individual questions, the crowd performed considerably better than individual respondents, with average performance in approximately the 85th percentile.

Finally, we examined the effect of social influence, randomly assigning participants to one of three different social conditions: (1) "concensus", in which participants saw the cumulative crowd response before providing their own answer; (2) "most recent", in which participants saw the three most recent answers; and (3) "most confident", in which participants saw three answers from the most confident individuals, based on self-reported assessments. For the latter two conditions—"most recent" and "most confident"—we found that crowd performance was qualitatively similar to the non-social, control condition. However, for the "consensus" condition, the crowd performed worse than when respondents did not receive any social signals. Notably, this consensus condition mirrors the design of many online rating sites, in which users can see the aggregate rating of others before providing their own rating. While such a design has value (e.g., it facilitates use by those who simply want to see the information, rather than providing a review themselves), our results suggest that it can also hurt the quality of results.

Related Work

The wisdom of the crowd effect

There is an extensive body of work documenting the wisdom-of-crowds phenomenon, including properties considered for it to be successful, as well as its limitations. While an exhaustive literature review is beyond the scope of this paper, we focus on those studies most closely related to ours.

Evidence of the phenomenon has been found in a wide range of domains: estimation tasks testing real-world knowledge regarding geographical facts and crime statistics (Lorenz et al. 2011), rank ordering problems (e.g., ranking U.S. presidents in chronological order) (Lee, Steyvers, and Miller 2014; Miller and Steyvers 2011), recollecting information from memory (Steyvers et al. 2009), and spatial reasoning tasks (Surowiecki 2005). But not all studies have been able to replicate this success. For example, Burnap et al. consider crowd evaluation of engineering design attributes and find that clusters of consistently wrong evaluators exist along with the cluster of experts. The authors conclude that both averaging evaluations and a crowd consensus model may not be adequate for engineering design tasks (Burnap et al. 2015).

This lack of consensus is also evident among the set of studies that consider prediction domains. In the context of predicting outcomes for competitive sporting tournaments, collective forecasts were found to consistently perform above chance and to be as accurate as predictions based on official rankings (Herzog and Hertwig 2011). In another study involving a competitive bidding task, Lee et al. considered eleven different methods to aggregate answers, and found that aggregation improves performance (Lee, Zhang, and Shi 2011). In contrast, in the betting context considered by Simmons et al., the authors found no evidence of a wisdom-of-crowds phenomenon. The authors attribute the failure to the fact that "most bettors have high intuitive confidence and are therefore quite reluctant to abandon it". Similarly, crowd predictions made by thousands of people competing in a fantasy football league were found to predict favorites in over 90% of the games, even though favorites and underdogs were equally likely to win against the spread (Simmons et al. 2010). These studies suggest that crowd wisdom may not prevail in contexts in which emotional, intuitive responses conflict with more rational, deliberative responses (Tversky and Kahneman 2000; Simmons et al. 2010).

Several studies focus on the question of how to best extract collective wisdom. Numerous studies have shown that simple aggregation techniques (e.g., using the mean or median for open-ended questions, or the majority vote for categorical questions) often perform just as well as more complex methods, including confidence-weighted aggregation, Bayesian methods, and the Thurstonian latent variable model (Miller and Steyvers 2011; Griffiths and Tenenbaum 2006; Prelec, Seung, and McCoy 2017; Budescu and Chen 2014; Hemmer, Steyvers, and Miller 2010). Simple aggregation, however, has often been found to perform reasonably well, if not on par with more complex models (Steyvers et al. 2009), across a variety of domains.

Effects of social influence

There is also mixed evidence for a wisdom-of-crowds effect in the presence of social influence. A series of studies have found that social influence can improve crowd estimates. Given that the information provided is accurate, there is evidence to suggest that it may improve crowd performance. Jayles et al. performed experiments in which subjects were asked to estimate quantities about which they had very little prior knowledge, before and after having received social information. Virtual "experts" providing the correct answer

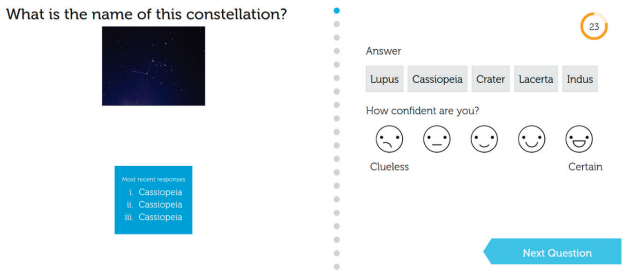

Figure 1: Sample categorical question showing the three most recent responses, self-assessed confidence prompt, and timer.

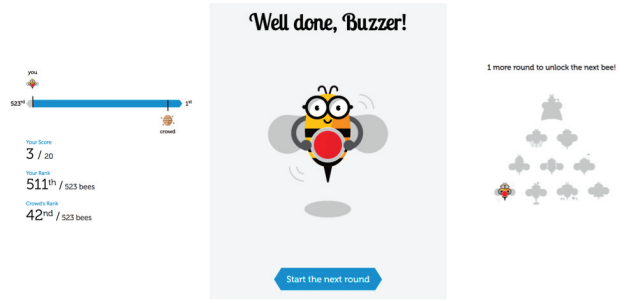

Figure 2: Example screenshot shown upon completion of a domain. The left panel shows the participant's rank in the domain relative to all other participants and to the crowd. The right panel shows the series of avatars the participant has yet to unlock as they progress through the game and complete domains.

for each question were inserted at random into the sequence of participants, and were found to help the group improve its collective accuracy (Jayles et al. 2017) .

A number of studies, however, have found that social influence may be beneficial even without this correctness constraint. For example, Miller et al. (Miller and Steyvers 2011) found that iterative communication between subjects on rank ordering tasks led to better estimates in reconstructing the correct answer compared to that of independent subjects. In a competitive gaming context of fantasy soccer, Goldstein et al. found that many players would do better by simply imitating the strategy of a player who has done well in the past, suggesting that social influence would be beneficial (Goldstein, McAfee, and Suri 2014).

Another set of studies provide a more nuanced view, showing that the type of social influence matters. In a "guess the number of sweets" task, King et al. find that individuals with access to the previous guess, mean guess, or a randomly chosen guess, tended to over-estimate the number of sweets, which undermined the crowd estimate. Providing the current best guess, however, prevented very large (inaccurate) guesses and resulted in convergence towards the true value and accurate crowd estimates (King et al. 2012).

Page et al. ran a controlled experiment where participants were randomly assigned to one of two network structures: centralized, where randomly selected participants placed in prominent positions to control information flow, or decentralized, in which everyone was equally influential. The authors tested crowd performance on a series of estimation tasks (34 groups of 40 people), and found that in centralized networks, the accuracy of the group depended entirely on the accuracy of a few influencers, while in decentralized networks, the average belief or opinion became more accurate after people communicated with each other (Page 2008).

In a separate study, Koriat considers a perceptual task and a general-information task, finding that group deliberation affected performance in the same direction, improving accuracy when individual accuracy was better than chance, but impairing it when individual accuracy was below chance. For consensually incorrect questions, group interaction impaired accuracy (Koriat 2015).

In contrast to this, are a number of studies that find evidence of social influence undermining the wisdom of crowd effect. For simple factual estimation tasks, Lorenz et al. find that subjects who could reconsider their response after having received average or full information of previous responses converged to worse answers (Lorenz et al. 2011). Knowledge about estimates of others was found to narrow the diversity of opinions, which undermined the wisdom of crowd effect by diminishing the diversity of the crowd without improvements of its collective error and shifting the position of the truth to peripheral regions of the range of estimates,

Muchnik et al. (Muchnik, Aral, and Taylor 2013) ran a large-scale randomized experiment on a Reddit-like website, finding that disclosing prior ratings created significant bias in individual rating behavior, leading to herding effects that were consequential to collective outcomes. In a cultural market setting, Salganik et. al. studied the effects of social influence on an online music platform where over 14,000 participants downloaded up to 48 previously unknown songs, either with or without knowledge of previous participants' choices. The authors found that increasing the strength of social influence increased both inequality and unpredictability of success. Success was also only partly determined by quality: the best songs rarely did poorly, and the worst rarely did well, but any other result was possible (Salganik, Dodds, and Watts 2006).

Experiment Design

To systematically investigate the wisdom of crowds phenomenon, and particularly the effects of social influence on collective judgment, we conducted a large-scale online experiment in which participants could answer up to 1,000 questions drawn from 50 topical domains. Each domain included questions on a specific topic, and was comprised of either 20 open-ended questions with numerical answers, or 20 categorical questions with categorical answers. Domains spanned four different types of media (text, image, video and audio) and included tests of explicit knowledge (e.g., factual questions, popular culture, spatial reasoning), tacit knowledge (e.g., emotional intelligence, foreign language skills), and prediction ability (e.g., election outcomes, box office

	Proportion
Female	50%
Male	49%
Decline to answer	1%
Age (18 – 24)	12%
Age (25 – 30)	26%
Age (31 – 40)	31%
Age (41 – 50)	16%
Age (51 – 60)	10%
Age (over 60)	4%
Some High school	6%
High school graduate	9%
Some college	32%
College graduate	38%
Some postgraduate work	4%
Post-graduate degree	10%

Table 1: Respondent characteristics, n=1,707.

success of upcoming movies). The full list of domains is listed in Table 2. In addition to asking respondents to answer the substantive questions, we elicited self-reported confidence from participants (on a 5-point scale) for each question.

To examine the effect of social influence, participants were randomly assigned to one of four different conditions in which they saw varying degrees of information on the responses of others: "consensus", "most recent", "most confident", and a control condition where respondents received no social information. The "most recent" condition displayed the previous three responses to the question. The "consensus" condition displayed the three most frequently selected responses in order from highest to lowest if the question was categorical, and the median answer up to that point if the question was open-ended. Finally, the "most confident" condition displayed the previous three responses with the highest self-reported confidence [1]. In all three social influence conditions, the first three participants to respond did not see any information about previous answers.

The experiment was run for three weeks on Amazon's Mechanical Turk. Domains were presented to participants in random order. Respondents were paid a flat amount of $0.40 for each completed domain. To incentivize accuracy, respondents received an additional bonus payment based on their ranking relative to others who completed the task. The bonus payment ranged from 0 to $0.20, with an average payment of $0.10 per completed domain. In total, respondents earned, on average, approximately $9 per hour.

We also incorporated several gamification aspects in order to encourage continued participation. In particular, respondents progressed through a series of avatars that they "unlocked" as they completed domains. A timer was also included for each question, which served both to discourage respondents from looking up information online and also to provide a timed objective to increase engagement. Figures 1 and 2 show screenshots of the online platform that was used to collect responses.

We asked respondents to answer questions on their own, without the aid of any outside materials. Although we do not have reason to believe that participants deviated from these instructions, we cannot be sure that it did not happen. If such behavior did occur, these individuals can be considered "experts" for our purposes.

Domain and question generation

A total of 50 domains were selected to cover a large variety of knowledge categories. Within each domain, the 20 questions were crowdsourced to a volunteer group of nearly 100 undergraduate students who were instructed to find an online corpus of questions for each domain and to then select 20 questions at random. For example, for the domain that asked participants to estimate the population of a country, the students compiled a list of all countries in the world, and selected a set of 20 at random. In effect, we thus used a crowd to help design and study the wisdom of crowds at scale. In 5 of the 50 domains, we asked participants to estimate the likelihood of a future event—like the election of a world leader or the winner of a sporting contest—on a scale from 0% to 100%. The "correct answer" for these domains was defined to be 100 if the event ultimately occurred, and 0 if it did not occur.

Before launching the experiment, the full corpus of questions and the selection strategy proposed by the undergraduate students was reviewed by the authors. Prior to running the full experiment, six small pilot tests were run on Mechanical Turk. The pilot tests helped us to ensure that the questions were clearly phrased, and of appropriate difficulty. In particular, we aimed to avoid questions for which there was no scope to develop expertise, or all respondents were expected to have expertise, as there is little room to observe a wisdom-of-crowds effect at these extremes. For example, asking U.S. respondents on what date the holiday of "July 4th" occurs is a valid question, but would not constitute a sensible choice as the answer is given in the question.

Stanford's IRB reviewed and approved our research project. Prior to beginning the experiment, we provided participants with an information page that explained the purpose of the study and the payment scheme, and emphasized that all data collected were de-identified. Participants had the option to cease answering questions at any point during the study without providing any reason. Participation was restricted to English-speaking respondents in the United States.

Measuring crowd performance

We analyze crowd performance at two levels of aggregation: the question-level and the domain-level (i.e., across a group of 20 questions on a specific topic). At the question-level, we define the crowd answer for open-ended questions to be the median of all responses; and for categorical questions, the crowd answer is defined to be the most popular response. In both cases, we measure the relative accuracy of crowd

[1] If there were more than three responses having the highest confidence level, three of these answers were chosen at random to be displayed.

Category	Question prompt	Type of media	Question type
Knowledge	What year was this building built in?	Image	Open-ended
Knowledge	In which year was this book published?	Image	Open-ended
Knowledge	In what year was the car manufactured ?	Image	Open-ended
Knowledge	In what year was this painting created?	Image	Open-ended
Knowledge	What is the population of [country name]?	Image	Open-ended
Knowledge	In what year did the [famous historical event] occur?	Text	Open-ended
Knowledge	What language is this?	Audio	Categorical
Knowledge	Various logic puzzles.	Image	Categorical
Knowledge	What is the species of this tree?	Image	Categorical
Knowledge	What is the name of this constellation?	Image	Categorical
Knowledge	Which country is the bill from?	Image	Categorical
Knowledge	Which country does this flag belong to?	Image	Categorical
Knowledge	Which country does this land border correspond to?	Image	Categorical
Knowledge	What is the name of this flower?	Image	Categorical
Knowledge	Which of the following is a synonym for [word]?	Text	Categorical
Knowledge	What is the breed of this dog?	Image	Categorical
Knowledge	In which language is the text is written?	Image	Categorical
Knowledge	Which pair of words has the same relationship as [X] : [Y]?	Text	Categorical
Knowledge	Who composed this [famous classical music piece]?	Audio	Categorical
Knowledge	What is the per-capita GDP of [name of country] in US dollars?	Text	Open-ended
Popular culture	In which year was this movie released?	Image	Open-ended
Popular culture	How many times will the following message be re-tweeted?	Image	Open-ended
Popular culture	How old is [celebrity name]?	Text	Open-ended
Popular culture	What does [common saying] mean?	Text	Categorical
Popular culture	Which artist/band interpreted this song?	Audio	Categorical
Popular culture	Which magazine published the headline?	Image	Categorical
Popular culture	In which of the following movies was this featured as a theme song?	Audio	Categorical
Tacit	How many calories does this [food item] contain?	Image	Open-ended
Tacit	Estimate the price in USD as listed on Amazon for the following product.	Image	Open-ended
Tacit	What is the average energy consumption of a typical [name of common appliance] in Watts? As a benchmark, a typical light bulb uses 60 - 100 Watts.	Image	Open-ended
Tacit	What emotion is being expressed in this image?	Image	Categorical
Tacit	What language is this?	Audio	Categorical
Tacit	In which direction will the ball go?	Video	Categorical
Tacit	Which of these [category] and [products] had the highest sales revenue in the U.S. in 2016?	Text	Categorical
Tacit	Will the following product be funded by Kickstarter?	Image	Categorical
Tacit	On what date in 2017 will [name of U.S./international holiday] fall?	Text	Categorical
Tacit	What musical instrument is this?	Audio	Categorical
Tacit	Was this US election news story real?	Image	Categorical
Tacit	Various questions related to negotiation skills, business ethics, and interview skills.	Text	Categorical
Tacit	Various questions regarding civil rights in the U.S. relating to privacy and police encounters.	Text	Categorical
Spatial reasoning	What is the distance in miles between [name of state, city in the U.S.], and [name of state, city in the U.S.]?	Text	Open-ended
Spatial reasoning	How many [country name] fit into the continental U.S.?	Image	Open-ended
Spatial reasoning	What is the weight of this object (in pounds)?	Image	Open-ended
Spatial reasoning	Under which cup is the ball located at the end of the trick?	Video	Categorical
Spatial reasoning	Various spatial reasoning puzzles.	Image	Categorical
Prediction	What is the likelihood that [political event] will occur before [date in 2017]?	Text	Open-ended
Prediction	What do you think the rating for this movie will be on Rotten Tomatoes?	Image	Open-ended
Prediction	What is the likelihood that [movie name] will win the Academy Award for [Academy award category]?	Image	Open-ended
Prediction	What is the likelihood that [technology / business event] will occur before [date in 2017]?	Text	Open-ended
Prediction	Which round will the [name of U.S. basketball team] make it to in the 2017 NCAA Tournament?	Text	Categorical

Table 2: Category, prompt, media, and question type for the 50 domains tested.

answers in terms of its percentile rank among the individual responses for that question. For example, a percentile rank of 70% means the crowd performed better than 70% of individuals who answered that question. Specifically, we first rank order the individual responses and the crowd answer by their distance from the ground truth. In the case of

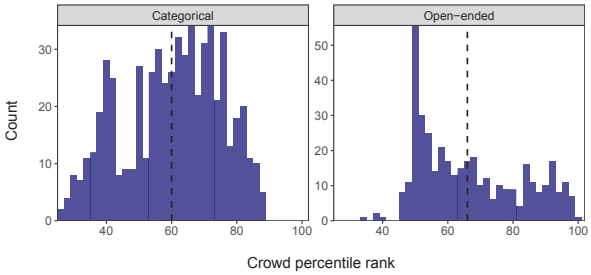

Figure 3: Crowd ranks for categorical and open-ended questions. The dotted lines represent the mean crowd rank (the 60^{th} and 66^{th} percentile for categorical and open-ended questions, respectively).

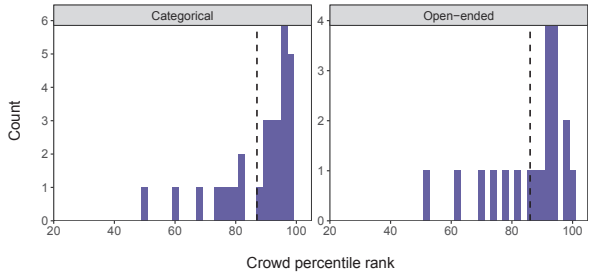

Figure 4: Crowd ranks for categorical and open-ended and domains. The dotted lines represent the mean crowd rank (the 87^{th} and 86^{th} percentile for categorical and open-ended domains, respectively).

open-ended questions, this distance is simply the absolute error of the response; and in the case of categorical questions, the distance is 0 if the response was correct and 1 otherwise. If there are ties—which occurs often for categorical questions—the rank of each tied entry is the average position for those ties in the list, which is the default behavior in `R`'s `rank` function. Finally, to convert from ranks to percentiles, we divide by $n + 1$, where n is the number of individual responses for that question, and we add 1 to account for the fact that the crowd answer is also in the ranked list.

The vast majority of studies on the wisdom-of-crowds effect have focused on question-level analysis. However, in many real-world contexts, groups are called upon for repeated assessments in a focused domain. Some examples include corporate boards, hedge fund managers, and academic review committees. To compute domain-level crowd performance, we first create an aggregate domain-level score for each respondent—where we treat the "crowd" as an additional respondent. For categorical questions, a respondent's domain-level score is simply the number of questions answered correctly. For open-ended questions, a respondent's domain-level score is that individual's average question-level rank. Then, as before, we define the relative performance of the crowd as its percentile rank among the domain-level scores for all respondents who completed that domain.

For our three social influence conditions, we likewise compute question-level and domain-level crowd performance. In this case, we compute the crowd answer based on the responses in the social condition, but in order to make consistent comparisons, we compute relative performance by benchmarking to the respondents in the *control* condition.

Results

We received approximately 510,000 responses from 1,707 respondents. On average, more than 100 individuals answered each of the 1,000 questions under each of the four experiment conditions (one control condition plus three social influence conditions). In total, 50% of participants were female, the median age was 37, and 84% of respondents reported having at least some college education (Table 1).

Question-level crowd performance

We start by considering question-level performance. We find that the average question-level crowd percentile rank for open-ended and categorical questions is 66 (s.e. 0.8) and 60 (s.e. 0.6), respectively. Thus, on a large and diverse corpus of questions, we find evidence that the crowd, on average, indeed outperforms the typical member of the constituent group.

There is, however, significant variation in crowd performance across questions, as shown in Figure 3. Whereas on some questions, the crowd achieves only a modest improvement over individual respondents, on others the crowd response achieves almost perfect performance, ranking above the 95th percentile. On some questions—particularly among the categorical questions—the crowd even ranks below 50%, apparently worse than the average member of the group. We note though, that this is in large part a statistical artifact of how ties are broken when computing crowd performance.

These results have two, somewhat different interpretations. On the one hand, our data support the conventional wisdom that crowds often perform better than the average member of the crowd. But, on the other hand, the amount of heterogeneity we see indicates that the wisdom-of-crowd effect is highly context dependent. This variation suggests that there is considerable nuance in when a wisdom-of-crowds effects holds, and helps to explain why past studies have not consistently found crowds to outperform individuals.

Domain-level crowd performance

We next consider domain-level performance, finding that the mean domain-level crowd percentile rank is 86 (s.e. 2.9) for open-ended domains and 87 (s.e. 2.2) for categorical domains. In particular, domain-level performance is considerably better than question-level performance—by more than 20 percentage points, on average. Further, as shown in Figure 4, domain-level performance is quite good in nearly all of the domains we consider. In Figure 5a, we directly compare question-level and domain-level performance for every domain. Specifically, for every domain, we compare the average question-level performance of the crowd (on the horizontal axis) to the domain-level performance (on the vertical axis). For every domain, there is a sizable improvement

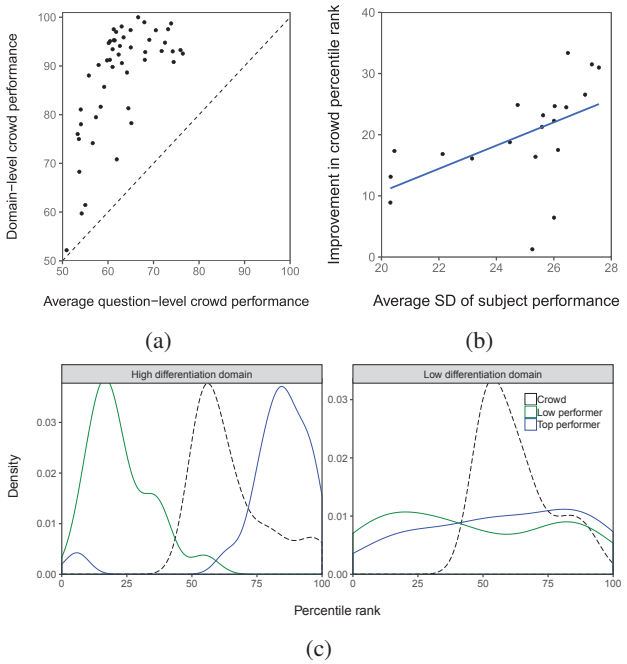

Figure 5: (a) Domain crowd ranks are greater than or equal to the average question-level crowd rank for constituent questions for every domain. (b) The improvement in crowd percentile rank is positively correlated to the amount of variability (standard deviation) in subject performance. (c) Distribution of ranks for a high performing respondent, a low performing respondent, and the crowd, for two sample domains. Both domains had an average question percentile rank of 62, but the crowd domain-rank improved by 32 percentage points in the low-differentiation domain (estimating retail prices), and only 13 percentage points in the high-differentiation domain (estimating country population).

when moving from individual questions to a domain-level aggregate. Thus, at the level of domains, we find that there is a large and consistent wisdom-of-crowds effect.

To better understand what drives this jump in performance for domains compared to individual questions, we placed domains on a spectrum that reflects differentiation in expertise. For our purposes, we quantified such differentiation by first computing the consistency of each individual respondent across questions in that domain. For example, if the typical respondent achieved similar performance across the full set of questions in a domain—meaning that some respondents consistently did well and others consistently did poorly—we considered that a "high differentiation" domain. One such domain was estimating the population of various countries, where, apparently, some respondents could do this task quite well and others could not. Conversely, if the typical respondent exhibited high variability in performance across questions, we considered that a "low differentiation" domain. As an example, we found little differentiation in performance when estimating the retail price of an item. At least in our pool of respondents, participants did not easily

	Question-level		Domain-level	
Social Condition	Open-ended	Discrete	Open-ended	Discrete
Control	66 (0.22)	60 (0.72)	86	87
Most recent	69 (0.17)	60 (0.72)	88	85
Most confident	67 (0.16)	59 (0.71)	81	83
Consensus	63 (0.19)	57 (0.67)	81	80

Table 3: Mean crowd percentile rank for open-ended and categorical domains (20 and 30 domains, respectively). Absolute crowd performance is given in parenthesis: average absolute relative errors are reported for open-ended questions, and the proportion of questions the crowd got correct are reported for discrete questions.

partition into "experts" and "non-experts" in this domain.

For these example "high differentiation" and "low differentiation" domains, Figure 5c graphically depicts the distribution of performance for low-performers (performing at the 20th percentile, in green), high-performers (performing at the 80th percentile, in blue), and the "crowd" (dashed line). Both domains had similar average question-level crowd rank of about 65%. But, importantly, in the high-differentiation domain (estimating population), the high-performing respondent does consistently better than the crowd; in the low-differentiation domain (estimating prices), that pattern does not hold. As a result, when aggregated to the domain-level, the crowd outperforms nearly all respondents in the low-differentiation domain, jumping 32 percentage points from question-level performance to domain-level performance. In the high-differentiation domain, there is a persistent subset of "experts" that the crowd cannot beat, and, accordingly, the question-level to domain-level jump is only 13 percentage points.

Figure 5b adds more quantitative detail to this pattern. On the horizontal axis, domains are ordered by average within-respondent standard deviation in performance, with low standard deviation (high differentiation) domains on the left, and high standard deviation (low differentiation) domains on the right. The plot confirms the intuition from our two examples above: the jump from question-level to domain-level performance (on the vertical axis) increases as one moves from high-differentiation to low-differentiation domains.

The effect of social influence on crowd performance

We conclude our analysis by investigating the wisdom-of-crowd effect in the presence of social influence. To recap, respondents were randomly assigned to one of three social conditions: "consensus", "most recent", or "most confident", or the control condition (in which there was no social influence).

Our results are summarized in Table 3, which displays the absolute performance and mean rank across all questions and domains by question-type, and in Figure 6, which shows the mean difference in performance between each social condition and control, averaged across all questions and domains [2].

[2]As an additional benchmark, the average probability of randomly guessing the correct answer on categorical questions is 0.25.

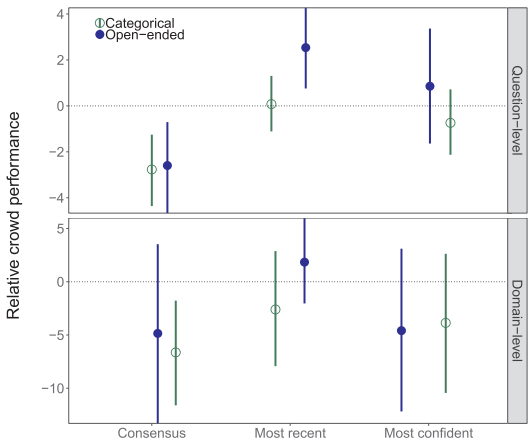

Figure 6: Question-level and domain-level crowd percentile ranks benchmarked to respondents in the control condition. The y-axis represents the difference between the social crowd and the control crowd, averaged across all questions. Error bars represent the 95% confidence interval of the average difference. Displaying the consensus answer leads to a significant decrease in average crowd performance for both discrete and open-ended questions.

Notably, the "consensus" condition exhibits *worse* performance than control, both for question-level and domain-level measures of performance, and for both open-ended and categorical questions. (In the case of open-ended questions evaluated at the domain level, the point estimate indicates that "consensus" is worse than "control", but the result is not statistically significant; in the other three combinations, the gap is statistically significant.) For the other two social conditions—"most recent" and "most confident"—we do not find statistically significant differences from the control condition. In the case of "most confident", we note that our results are in line with previous work that finds that methods which incorporate self-reported confidence do not lead to improvement in group estimation (Madirolas and de Polavieja 2015).

Why does seeing the "consensus" answer degrade performance? Our data suggest that it is because respondents heavily anchor to the "consensus" response, prompting vicious cycles in which initially inaccurate responses can pull down the entire crowd. To see this, we partition questions into those that had "accurate" starts and those that had "inaccurate" starts, based on the initial three responses. The first three respondents never saw any social cues, so those that happened to start in a worse position did so by chance alone. As shown in Figure 7, the crowd rank for questions with inaccurate starts (i.e., those for which the median or modal answer of the first three respondents ranks in the bottom 50th percentile) does not rebound in the consensus condition. This anchoring effect occurs even though participants are shown the total number of responses on which the consensus is based. However, in the other two social conditions, the crowd seems able to recover, appropriately ignoring initial inaccuracies. This result suggests that care should

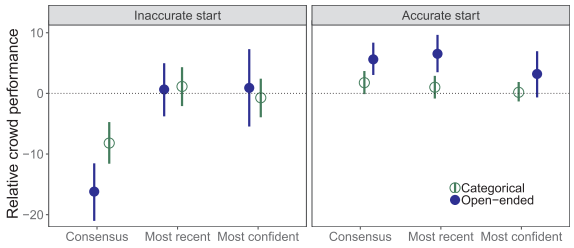

Figure 7: Cumulative crowd rank, grouped by initial performance. Initial performance is measured by the percentile rank of the median of the first three responses. An inaccurate start indicates a percentile rank of 50 or below. The consensus condition does not rebound from poor initial starting positions.

be taken when incorporating social influence into platforms that leverage the wisdom of crowds, as an initial "madness" could have long-lasting consequences.

Conclusion

In one of the largest experiments to date on the wisdom-of-crowds effect—involving 1,000 questions, nearly 2,000 participants, and over 500,000 responses—-our results paint a nuanced picture of the phenomenon. When analyzing performance at the level of individual questions, as is standard in the literature, we find the crowd, on average, outperforms its constituent members. But there is also substantial variation across questions—even across questions within a single domain—indicating that the wisdom-of-crowds effect is sensitive to the exact context. However, when we aggregate to the level of domains, the crowd quite consistently outperforms individuals, often by a large margin. This difference between question-level and domain-level performance appears to stem from the fact that even "expert" respondents do not always perform well. The consistency of the crowd leads to cumulative advantages when performance is measured on an extended battery of questions. Finally, we examined the effect of social influence on crowd performance. Showing social cues related to recent or confident answers does not appear to qualitatively affect our results. But showing respondents the crowd's current consensus can trigger cascades, in which initial inaccuracies persist, degrading overall performance.

At least since Galton's *vox populi* over a century ago, there has been enduring interest and investigation into the power of collective judgments. To this expansive literature—which has applied a wide variety of analytic methods to study a diverse set of domains and populations—we have attempted to bring a degree of consistency, testing performance on a large corpus of questions in a uniform manner and on a fixed population. Our approach has, we believe, helped us shed new insights on an old phenomenon, though it also leaves many questions unanswered. Among those are what domain characteristics explain the variance in crowd performance, strategies for anticipating which domains may be amenable to leveraging collective intelligence, and de-

veloping other crowd selection and aggregation approaches that might improve crowd performance in these domains. We hope, though, that our work provides firmer footing for future researchers to continue investigating the wisdom of crowds.

Data availability

The data collected during the experiment, along with any media provided for each question are available at https://github.com/stanford-policylab/wisdom-of-crowds.

Acknowledgments

We thank Rajan Vaish for helpful discussion and comments as well as for helping us lead the group of undergraduate researchers who contributed to the question corpus. The undergraduate researchers were recruited as part of the Crowd Research Initiative (Vaish et al. 2017), and we are grateful to Imanol Arrieta Ibarra and Michael Bernstein for helping us lead the research group. Finally, we thank the undergraduate students that helped us compile the corpus of questions, as listed in (Mysore 2015), as well as Nikhil Prakash for helpful feedback.

References

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

Burnap, A.; Ren, Y.; Gerth, R.; Papazoglou, G.; Gonzalez, R.; and Papalambros, P. Y. 2015. When crowdsourcing fails: A study of expertise on crowdsourced design evaluation. *Journal of Mechanical Design* 137(3):031101.

Faria, J. J.; Dyer, J. R.; Tosh, C. R.; and Krause, J. 2010. Leadership and social information use in human crowds. *Animal Behaviour* 79(4):895–901.

Galton, F. 1907. Vox populi (the wisdom of crowds). *Nature* 75:450–451.

Goel, S.; Reeves, D. M.; Watts, D. J.; and Pennock, D. M. 2010. Prediction without markets. In *Proceedings of the 11th ACM conference on Electronic commerce*, 357–366. ACM.

Goldstein, D. G.; McAfee, R. P.; and Suri, S. 2014. The wisdom of smaller, smarter crowds. In *Proceedings of the fifteenth ACM conference on Economics and computation*, 471–488. ACM.

Griffiths, T. L., and Tenenbaum, J. B. 2006. Optimal predictions in everyday cognition. *Psychological science* 17(9):767–773.

Hemmer, P.; Steyvers, M.; and Miller, B. 2010. The wisdom of crowds with informative priors. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.

Herzog, S. M., and Hertwig, R. 2011. The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition. *Judgment and Decision Making* 6(1):58–72.

Hill, S., and Ready-Campbell, N. 2011. Expert stock picker: the wisdom of (experts in) crowds. *International Journal of Electronic Commerce* 15(3):73–102.

Hueffer, K.; Fonseca, M. A.; Leiserowitz, A.; and Taylor, K. M. 2013. The wisdom of crowds: Predicting a weather and climate-related event. *Judgment and Decision Making* 8(2):91–105.

Jayles, B.; Kim, H.-r.; Escobedo, R.; Cezera, S.; Blanchet, A.; Kameda, T.; Sire, C.; and Theraulaz, G. 2017. How social information can improve estimation accuracy in human groups. *Proceedings of the National Academy of Sciences* 114(47):12620–12625.

Kaplan, A.; Skogstad, A.; and Girshick, M. A. 1950. The prediction of social and technological events. *Public Opinion Quarterly* 14(1):93–110.

King, A. J.; Cheng, L.; Starke, S. D.; and Myatt, J. P. 2012. Is the true 'wisdom of the crowd' to copy successful individuals? *Biology Letters* 8(2):197–200.

Koriat, A. 2015. When two heads are better than one and when they can be worse: The amplification hypothesis. *Journal of Experimental Psychology: General* 144(5):934.

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

Lee, M. D.; Zhang, S.; and Shi, J. 2011. The wisdom of the crowd playing the price is right. *Memory & cognition* 39(5):914–923.

Liu, Y.; Chen, F.; Kong, W.; Yu, H.; Zhang, M.; Ma, S.; and Ru, L. 2012. Identifying web spam with the wisdom of the crowds. *ACM Transactions on the Web (TWEB)* 6(1):2.

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

Mackay, C. 1841. *Memoirs of extraordinary popular delusions and the madness of crowds*. George Routledge and Sons.

Madirolas, G., and de Polavieja, G. G. 2015. Improving collective estimations using resistance to social influence. *PLoS computational biology* 11(11):e1004594.

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

Moore, T., and Clayton, R. 2008. Evaluating the wisdom of crowds in assessing phishing websites. In *Financial Cryptography and Data Security*. Springer. 16–30.

Muchnik, L.; Aral, S.; and Taylor, S. J. 2013. Social influence bias: A randomized experiment. *Science* 341(6146):647–651.

Mysore, S. 2015. Investigating the "wisdom of crowds" at scale. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, UIST '15 Adjunct, 75–76. New York, NY, USA: ACM.

Page, S. E. 2008. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies-New Edition*. Princeton University Press.

Prelec, D.; Seung, H. S.; and McCoy, J. 2017. A solution to the single-question crowd wisdom problem. *Nature* 541(7638):532.

Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311(5762):854–856.

Simmons, J. P.; Nelson, L. D.; Galak, J.; and Frederick, S. 2010. Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research* 38(1):1–15.

Steyvers, M.; Miller, B.; Hemmer, P.; and Lee, M. D. 2009. The wisdom of crowds in the recollection of order information. In *Advances in neural information processing systems*, 1785–1793.

Surowiecki, J. 2005. *The wisdom of crowds*. Anchor.

Tversky, A., and Kahneman, D. 2000. *Choices, values, and frames*. Cambridge University Press.

Vaish, R.; Gaikwad, S. N. S.; Kovacs, G.; Veit, A.; Krishna, R.; Arrieta Ibarra, I.; Simoiu, C.; Wilber, M.; Belongie, S.; Goel, S.; et al. 2017. Crowd research: Open and scalable university laboratories. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, 829–843. ACM.