Computational Brain & Behavior
https://doi.org/10.1007/s42113-018-0006-4

Making a Wiser Crowd: Benefits of Individual Metacognitive Control on Crowd Performance

Stephen T. Bennett[1] · Aaron S. Benjamin[2] · Percy K. Mistry[1] · Mark Steyvers[1]

© Springer International Publishing 2018

Abstract

The wisdom of the crowd refers to the finding that judgments aggregated over individuals are typically more accurate than the average individual's judgment. Here, we examine the potential for improving crowd judgments by allowing individuals to choose which of a set of queries to respond to. If individuals' metacognitive assessments of what they know is accurate, allowing individuals to opt in to questions of interest or expertise has the potential to create a more informed knowledge base over which to aggregate. This prediction was confirmed: crowds composed of volunteered judgments were more accurate than crowds composed of forced judgments. Overall, allowing individuals to use private metacognitive knowledge holds much promise in enhancing judgments, including those of the crowd.

Keywords Wisdom of the crowd · Self-direction · Metacognition

The earliest and most famous example demonstrating the wisdom of the crowd comes from a report by Galton (1907). In that example, nearly 800 visitors to an agricultural exhibition in England entered a contest in which they guessed the weight of an ox. The central finding, known broadly throughout the social sciences today, is that the average judgment of the group was impressively accurate—in fact, the median judgment came within 1% of the correct answer. The superior accuracy of such crowd judgments is evident in a wide variety of tasks, including complex combinatorial problems (Yi et al. 2012), recitation of lists in an appropriate order (Steyvers et al. 2009), and in predicting events with as-yet unknown outcomes (Lee and Danileiko 2014; Turner et al. 2014; Merkle et al. 2016; Mellers et al. 2014).

Despite a century of research on this important topic, an aspect of the original example from Galton has gone unappreciated. The fairgoers in his data set were a self-selected bunch: they *chose* to provide a weight estimate. Not only that, they paid (a sixpenny) for the privilege of doing so (and to have the opportunity to win a prize). This may seem like a small matter, but there are reasons to think it might not be. Research on *metacognition* reveals that people are good judges of their knowledge and of the accuracy of their judgments. Allowing individuals to opt in to a particular judgment based on an assessment of their own expertise may in fact have created a crowd of exceptional wisdom in Galton's case. This is the question we pursue here: does allowing an individual the choice of when to respond improve the accuracy of the resultant crowd? There are both theoretical and practical reasons to care about this problem.

On the practical side, there are now a large number of crowd-sourcing platforms in which individuals choose which tasks to participate in. For example, prediction markets (e.g., Predictit, Tradesports), swarm intelligences (e.g., UNU), and forecasting tournaments like the Good Judgment Project (Mellers et al. 2014) all cede control to the individuals as to what tasks to perform. Recent analyses have shown that the choice of forecasting problems in the Good Judgment Project is related to forecasting skill (Merkle et al. 2017). These results suggest that the specific problems selected by individuals can provide valuable information about the person. However, it is not known whether the self-selection procedure reduces or enhances the wisdom of the crowd.

On the theoretical side, the results from the experiments reported here have the potential to inform our theories about

✉ Stephen T. Bennett
stbennet@uci.edu

[1] University of California, Irvine, USA

[2] University of Illinois at Urbana-Champaign, Illinois, USA

Published online: 11 July 2018

the origin of the benefit that arises from aggregating within a crowd. The benefit of the crowd has two general classes of explanations. The first is that knowledge for a particular question is diffusely distributed across the population; random perturbations from the truth that occur within a group cancel out in a large enough sample (Surowiecki 2004). By this reasoning, individuals are exchangeable since perturbations among individuals are random. In a second class of explanation, knowledge for a particular question is concentrated within a subset of individuals in the group, and others contribute little more than random noise. Averaging reduces the influence of random responders, leaving the signal from knowledgeable responders to reveal itself more clearly. By this explanation, individual differences in knowledge are paramount, and the selection of responders to contribute to a given query or task should depend heavily—perhaps exclusively—on those individual differences.

One way of ensuring that contributions come from knowledgeable sources is to seek out populations with particular expertise. Techniques have been developed that identify expertise or upweight expert judgments on the basis of calibration questions (Bedford and Cooke 2001), performance weighting (Budescu and Chen 2014), coherence and consistency weighting (Olson and Karvetsi 2013; Weiss et al. 2009), and consensus models (Lee et al. 2012). For an overview of these methods, see Steyvers and Miller (2015). Though these techniques each have their own advantages, there are also a number of challenges. How do we identify individuals with particular expertise, or domains of particular expertise within an individual? What if those individuals are difficult to find or costly to obtain? What if the domain under investigation is one that requires wide-ranging expertise?

The alternative approach reviewed here allows individuals to make their own judgments about their ability to contribute to the problem under investigation. There is ample reason to believe that such judgments are likely to be highly accurate. People successfully withhold responses in which they have low confidence, increasing the accuracy of volunteered responses. They can also vary the grain size of their answer, answering with great precision when they know their knowledge to be accurate and with lesser precision when they are unsure (Goldsmith and Koriat 2007). Control over selection of items for restudy (Kornell and Metcalfe 2006) and over the allocation of study time (Tullis and Benjamin 2011) benefits learners. All of these findings point to the skill with which individuals exert metacognitive control over their learning and remembering, and how that control benefits performance (Fiechter et al. 2016; Benjamin 2008a; Benjamin and Ross 2008b).

The specific choice about when and when not to respond to a query is helpful in an impressive variety of situations. In psychometrics, test-takers prefer the ability to choose which questions they are graded on Rocklin (1994), Schraw et al. (1998), and improve their performance by doing so. Psychometric models of this type of choice have demonstrated benefits in estimating subject characteristics (Culpepper and Balamuta 2017). Even non-human animals have the metacognitive ability to choose when to bet on their success in a particular trial (Kepecs and Mainen 2012; Middlebrooks and Sommer 2011). As noted above, the freedom to choose the grain size of reported memories substantially improves the accuracy of memories that are reported (Goldsmith and Koriat 2007; Koriat and Goldsmith 1996). These strategies all reflect the positive contribution of metacognitive processes and the benefits of permitting participants to self-regulate responding, but no research has yet examined the potential of self-regulation for improving crowd accuracy.

The allowance of individual metacognitive control can affect crowd cognition in complex and perhaps unanticipated ways. Responders' choices about when to respond based on their own knowledge impact both the quality and distribution of responses. Even if responders make good metacognitive choices that improve the quality of their responses, giving people this freedom may result in a shift in the distribution of responses wherein subsets of questions go unanswered. The consequences of unanswered questions may be high, and the net cost of this unanswered subset may outweigh benefits gained on other questions. Taken together, it is unclear how providing a group the ability to self-select questions will impact crowd response over a large set of questions or predictions, and with different performance metrics.

We investigated the effect of allowing responders to opt in to questions of their own choosing on the wisdom of the crowd effect in two experiments. In each experiment, one group of responders chose among a subset of binary-choice trivia questions and a control group answered randomly assigned questions. If responders use their metacognitive knowledge judiciously in service of selecting which questions to answer, we should see an advantage for crowds composed of *self-directed* responders over a typical (control) crowd. In Experiment 1, we matched participants in the self-directed condition with participants in the control condition in terms of total number of judgments and assessed performance on a set of relatively easy (Experiment 1a) and difficult (Experiment 1b) questions. In Experiment 2, we ceded further control to participants by allowing them to choose as many or as few questions as they wished from among a relatively easy set of questions.

Experiment 1

Method

Participants

166 participants were recruited through Amazon Mechanical Turk (AMT). Each participant was compensated \$1 for the 30 min the experiment was expected to take. Each participant was randomly assigned to Experiment 1a ($N = 83$; easy questions) or 1b ($N = 83$; difficult questions). In both experiments, each participant was randomly assigned to the opt-in (or "self-directed") condition, in which they had the ability to choose which questions to answer ($N = 39$ for both experiments) or to the control condition, in which questions were randomly assigned to them ($N = 44$ for both experiments). No participant completed more than one condition.

Stimuli

Stimuli consisted of 144 general knowledge binary-choice questions. The questions were drawn from 12 general topics: World Facts, World History, Sports, Earth Sciences, Physical Sciences, Life Sciences, Psychology, Space & Universe, Math & Logic, Climate Change, Physical Geography, and Vocabulary. In order to empirically determine how difficult the questions were, we conducted a pilot experiment in which 54 participants answered 48 questions each. Average accuracy across all 144 questions was 55.2%. We formed two sets of 100 questions each based on the easiest and most difficult questions, which resulted in an overlap of 56 questions between the easy and difficult question sets. The 100 easiest questions, which yielded 73% accuracy, comprised the stimulus set for Experiment 1a. The 100 hardest questions, which yielded 48% accuracy, comprised the stimulus set for Experiment 1b. Four example questions are shown in Table 1. No participants who completed the pilot study were recruited for Experiment 1.

Design and Procedure

Participants could view the survey description on AMT. If they selected the survey, they were redirected to another website (hosted using the Qualtrics platform). They were first directed to a study information sheet that provided details of the survey and compensation. If they agreed to continue, they answered demographic questions, and then were randomly assigned to the experiment and condition.

Participants were not aware of the existence of other conditions. Each participant viewed questions in five blocks of 20 questions each. They were instructed to rate the difficulty of each question from 1 (Very Easy) to 4 (Neither Easy nor Difficult) to 7 (Very Difficult). Then, if they were assigned to the self-directed condition, they were instructed to choose five questions to answer in that block. The participants in the control condition were randomly assigned five questions to answer. After rating the difficulty of all 100 questions and answering 25 total questions, participants were thanked for their time and given instructions on how to receive payment.

Scoring Crowd Performance

We utilized three general methods for measuring crowd performance. The first measure is based on the accuracy of the majority answer, which we term *crowd accuracy*. For each individual question, we score the crowd as correct (1) if the majority of the participants in a crowd answered correctly and incorrect (0) if the majority of the participants in a crowd answered incorrectly. Questions that elicited an equal number of correct and incorrect responses were assigned a value of 0.5. Unanswered questions were also assigned an accuracy value of 0.5, corresponding to the chance value of answering the question accurately with no knowledge. Crowd accuracy is then based on the average score across questions. We report the accuracy measure because it is easily interpretable and is widely used in this

Table 1 Example questions

Difficulty	Example
Hard	(1) The sun and the planets in our solar system all rotate in the same direction because: *(a) they were all formed from the same spinning nebular cloud*, or (b) of the way the gravitational forces of the Sun and the planets interact
	(2) The highest man-made temperature has been: (a) less than 1 million °C, or *(b) greater than 1 million °C?*
Easy	(1) Greenhouse effect refers to: *(a) gases in the atmosphere that trap heat*, or (b) impact to the Earth's ozone layer
	(2) Which is Earths largest continent by surface size? (a) North America, or *(b) Asia*

Correct answers are italicized

literature. However, it has low statistical power because the underlying observations are (mostly) binary.

The second measure, *proportion correct*, is based on a more fine-grained assessment of crowd performance. For each individual question, we assessed the proportion of respondents in the crowd that answered correctly. The overall proportion correct measure is based on the mean of these proportions. The third measure, *proportion better*, assesses the proportion of questions for which the opt-in condition outperformed the control condition (ignoring questions with equal accuracy between conditions).

The last two measures can detect differences in crowd performance even when the majority rule leads to the same answer. For example, if the opt-in and control condition reveal 80 and 70% correct response rates respectively (for each individual question), the crowd accuracy measure based on majority rule would not be able to distinguish between the two conditions, whereas the proportion correct and proportion better measures would reveal the advantage of the opt-in condition.

Results

Data from all experiments reported in this article are publicly available on the Open Science Framework (https://osf.io/nhv3s). For all of our analyses, we utilize Bayes factors (BFs) to determine the extent to which the observed data adjust our belief in the alternative and null hypotheses. There are numerous advantages of BFs over conventional methods that rely on p values (Rouder et al. 2009; Jarosz and Wiley 2014; Wagenmakers 2007), including the ability to detect evidence in favor of a null hypothesis and a straightforward interpretation. In order to compute the BFs, we used the software package JASP (Love et al. 2018) and a Bayes factor calculator available online (Rouder et al. 2009; Rouder 2018). In both cases, we maintained the default priors that came with the software when performing computations.

In our notation, $\mathrm{BF} > 1$ indicates support of the alternative hypothesis while $\mathrm{BF} < 1$ indicates support of the null hypothesis. For instance, $\mathrm{BF} = 5$ means the data are five times more likely under the alternative hypothesis than the null hypothesis. Similarly, $\mathrm{BF} = 0.2$ corresponds to an equal amount of support of the null hypothesis. When discussing BFs, we use the language suggested by Jeffreys (1961). In order to improve readability, BFs larger than 100,000 are reported as $\mathrm{BF} > 100,000$.

Raw Data

Figure 1 shows the full pattern of chosen and assigned questions across respondents in the self-directed and control condition as well as the correctness of individual answers (see Fig. 4 in the Appendix for the distribution of responses in the hard condition). The distributions reveal that some questions are chosen much more often than others. Note that

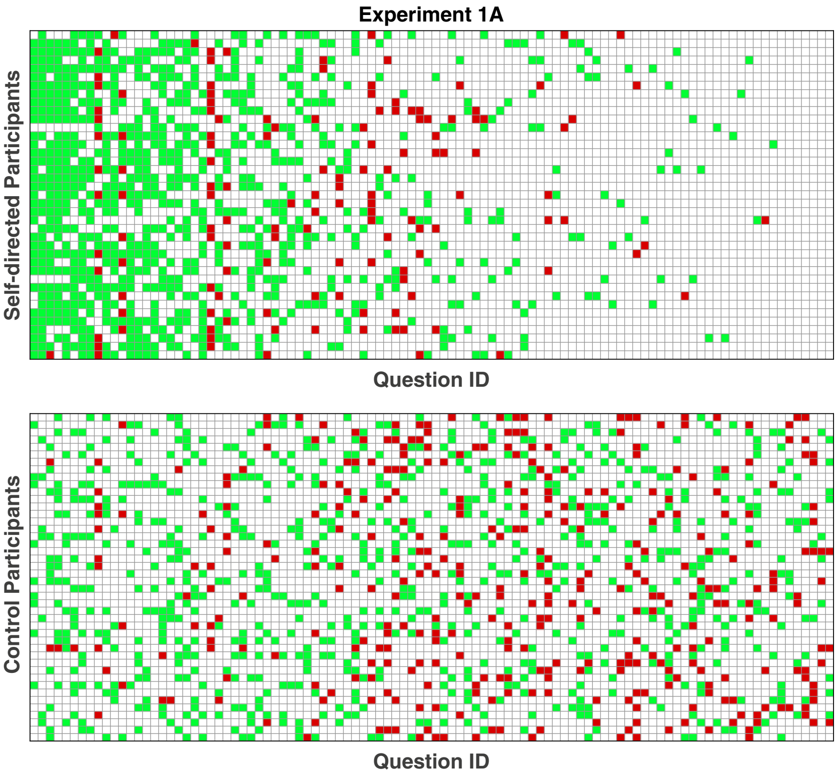

Fig. 1 Question responses for the self-directed and control participants in the easy condition (Experiment 1a) with questions sorted by the number of participants who selected the question in the self-directed condition. Green squares represent correct responses, red squares represent incorrect responses, and white squares represent no response. Self-directed participants tend to cluster around the same collection of questions when compared to control participants

Table 2 Average individual performance across conditions

Experiment	Control (%)	Opt-in (%)	Full opt-in (%)	BF
1a	67.27	86.05		> 100,000
1b	48.64	53.85		0.540
2	69.21	83.80		13,046
	69.21		82.75	7,581
		83.80	82.75	0.259

Each Bayes factor (BF) compares individual performance of the opt-in condition with the control condition within that experiment

in the self-directed condition, there were seven questions that no participant chose to answer in Experiment 1a and one such question in Experiment 1b. For the control condition, each question was randomly assigned to at least four participants in the control condition and therefore no question went unanswered.

Individual Accuracy Differences

First, we confirmed our assumption about the difficulty of questions in each condition. Participants in Experiment 1a (with easy questions) averaged 76.10% accuracy and those in Experiment 1b (with difficult questions) averaged 45.16%. In addition, Table 2 shows the average accuracy of individuals across conditions. We used a Bayesian *t* test to assess whether individual accuracy was higher or lower in the opt-in condition. In Experiment 1a, there is evidence that participants who opted in to questions exhibited higher average accuracy than those who were randomly assigned to questions. However, the data in Experiment 1b were ambiguous, providing little evidence one way or the other in terms of the relative accuracy of the opt-in and control conditions.

Crowd Performance

Table 3 shows the crowd performance under the three performance metrics introduced earlier, and summarizes the results of our analyses. In general, we found that crowds composed of self-selected judgments outperformed those with judgments from participants randomly assigned to questions. Analyses comparing crowd accuracy and proportion correct used a two-tailed Bayesian paired sample *t* test. To analyze proportion better, we used a Bayesian binomial test to assess if the rate of proportion better exceeded 50%.

For the easy questions (Experiment 1a), we found consistent evidence for a benefit of self-directed crowds over control crowds. While the evidence was only anecdotal for crowd accuracy, the more fine-grained measures of proportion correct and proportion better both provide decisive evidence that opting in benefits aggregate performance. For the hard questions (Experiment 1b), we found the same effect but with less decisive evidence. Specifically, we found moderate evidence from the crowd accuracy and proportion correct metrics that self-direction was beneficial to crowd performance. The weaker evidence for the harder questions is likely related to the ambiguous finding when comparing individual accuracy between the opt-in and control conditions. Taken together, these analyses demonstrate that crowds composed of responders who voluntarily opt in to questions are indeed superior.

Difficulty Ratings

Why is it that participants were more accurate in the self-directed condition? Presumably participants are choosing questions that they find easy. In accordance with this

Table 3 Crowd performance across conditions and the performance metrics crowd accuracy, proportion correct and proportion better

	Crowd accuracy				Proportion correct					
Experiment	Control (%)	Opt-in (%)	Full opt-in (%)	BF	Control (%)	Opt-in (%)	Full opt-in (%)	BF	Prop. better (%)	BF
1a	73.0	82.5		1.904	67.36	79.10		4847	73.49	1665
1b	46.0	58.5		4.031	49.48	56.91		3.983	60.82	1.219
2	78.5	76.0		0.1326	68.58	73.82		0.61	64.77	6.248
	78.5		82.0	0.1506	68.58		76.07	71.65	73.63	4504

Each Bayes factor (BF) compares performance of the opt-in condition with the control condition within that experiment hypothesis, we found a strong correlation between the probability of opting in to a question and its average difficulty rating in both Experiment 1a ($N = 100$, $r = -0.90$, BF $> 100,000$) and 1b ($N = 100$, $r = -0.88$, BF $> 100,000$).

We also investigated whether participants preferred questions that they rated as easier than their peers. We first identified the set of questions for each participant that were judged to be easier than the average rating. For this set of question-participant pairs, we computed the probability that the participant opted in to that question. For Experiment 1a and 1b, these probabilities were 38.43% ($N = 1996$) and 39.30% ($N = 1901$) respectively. Comparing these values to chance (25%) gave decisive evidence that participants chose questions that they rated as easier than their peers (BF $> 100,000$ for both Experiment 1a and b).

Overall, people are choosing those questions that are easier *for them*. This finding implies that there is a common metacognitive process by which people choose which questions to answer, based in large part on their metacognitive assessments of item difficulty and their unique expertise.

Simulating Opt-in Crowds Based on Rated Difficulty

With the current data set, it is not possible to directly assess how participants would perform on questions that they did not choose. However, we can simulate an opt-in decision for participants in the control condition based on their difficulty ratings (participants rated all questions). Previously, it was found that crowds composed of confident responses led to higher accuracy than same-sized crowds composed at random (Mannes et al. 2014).

We investigated the crowd performance for answers that were perceived to be below some threshold level of difficulty. For example, for a threshold of "4", we identified all participants, in the control condition only, who answered that question and rated the difficulty at or below "4." For comparison, we also composed a control crowd of equal size by randomly sampling any participant who answered that question ignoring the rated difficulty. Figure 2 shows crowd performance as assessed by the proportion correct metric for the simulated opt-in and random comparison groups. The results show that simulating opt-in in this way can yield better performing crowds when compared to randomly composed crowds of the same size. In addition, smaller crowds that include only answers from people who rated the question as easy outperformed the full crowd composed of all answers (corresponding to a cutoff of 7).

Discussion

In Experiment 1, we found that crowds formed from participants with the opportunity to self-select questions outperformed crowds that were formed from participants randomly assigned to questions. The evidence in support of this claim was decisive for the easier set of questions in Experiment 1a and substantial for the harder questions in Experiment 1b. Additionally, we observed that there appears to be a metacognitive process that governs the relationships among all of the observed behaviors. People select questions that are easy for them and then perform well on them when given the opportunity to answer them selectively. Simulating choice with these difficulty ratings improved crowd performance relative to random samples of crowds of similar size and the complete crowd.

Experiment 2: Full Choice

In Experiment 1, we demonstrated that a wiser crowd can be created by allowing responders to decide when they want to provide a response. In that experiment, we allowed participants to choose *which* questions to answer, and matched that group with a control group in terms of *how many* questions they answered. This methodological

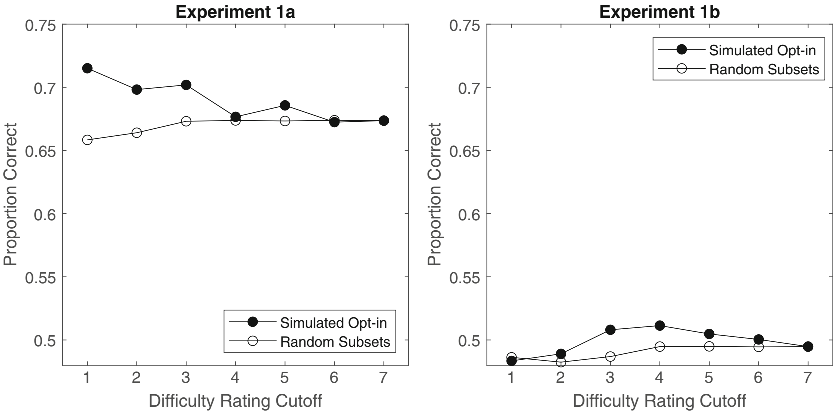

Fig. 2 Crowd performance for subsets of judgments below a difficulty rating (simulated opt-in) and randomly chosen judgments (random subsets)

choice had the benefit of ensuring a reasonable crowd size for most questions. However, since we observed a benefit to crowd performance from permitting *some* self-direction, a natural question is to ask whether or not *more* control over responding is even better. The specific additional freedom we grant participants in this experiment is to respond to as many questions as they desire. If question choice is driven by knowledge, then participants who have substantially more knowledge than others will now have the opportunity to contribute to a greater extent. Similarly, participants who have a relatively shallow pool of knowledge will be able to avoid answering questions for which they lack relevant knowledge.

Method

The stimuli and design were the same as Experiment 1a, with one additional condition. The new *full opt-in* condition allowed participants to choose to answer as many or as few questions as they wish. As in Experiment 1, participants chose questions, provided difficulty ratings for each question, and then answered the questions that they chose. To contrast with full opt-in, we now term what had been the self-directed condition in Experiment 1 the *partial opt-in* condition.

Participants A total of 118 participants were recruited through Amazon Mechanical Turk (AMT). Each participant was compensated $1 for the 30 min the experiment was expected to take. No participant completed more than one condition and no participants who completed the pilot study or Experiment 1 were recruited for this experiment. Participants were randomly assigned to the partial opt-in ($N = 39$), full opt-in ($N = 36$), or control ($N = 43$) condition.

Stimuli

The stimuli were the same as those used in Experiment 1a.

Design and Procedure

The procedure was the same as Experiment 1, with the exception of the new, full opt-in condition in which participants were instructed to respond to as many questions as they "felt they could answer well."

Analysis

We utilized the same three methods for measuring and comparing crowd performance as in Experiment 1.

Results

Responses from the full opt-in condition are shown in Fig. 3 and those from the partial opt-in condition can be viewed in Fig. 5 in the Appendix. As in Experiment

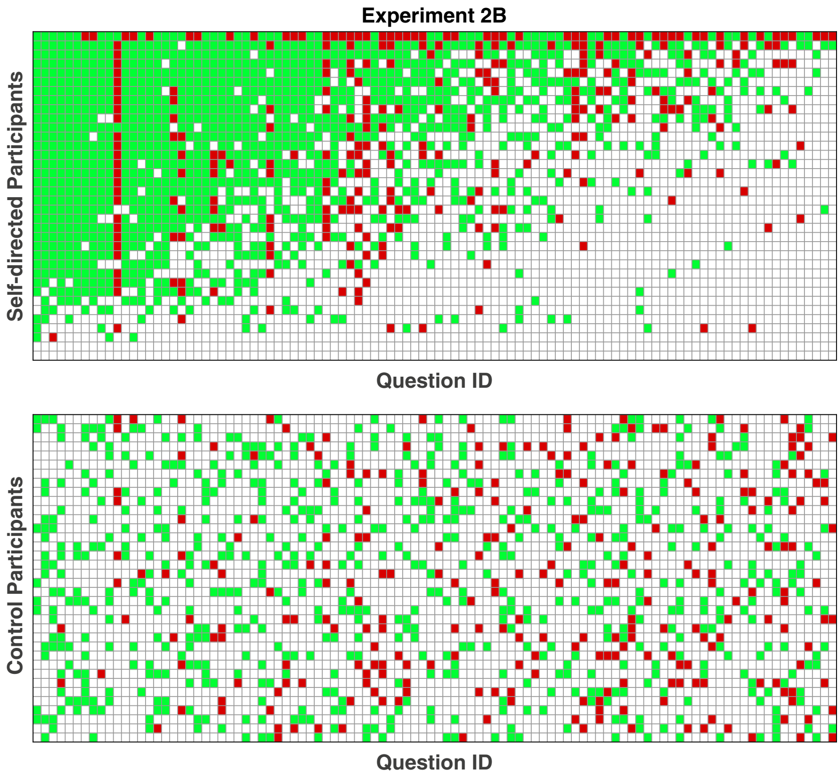

Fig. 3 Question responses for the self-directed participants in the full-choice condition as well as the control participants in Experiment 2. Questions are sorted by the number of participants who selected the question in the opt-in condition and self-directed participants are sorted by the number of questions they chose to answer. Green squares represent correct responses, red squares represent incorrect responses, and white squares represent no response. Self-directed participants tend to cluster around the same collection of questions when compared to control participants

1a, seven questions went unanswered in the partial opt-in condition. No question went unanswered in the full opt-in or control conditions. Participants in the full opt-in condition selected 44.11 (SD=25.22) of 100 questions on average, significantly more than the 25 questions per participant required in the other conditions ($t(35) = 4.546$, BF = 381.3). Table 2 shows that there is evidence that partial or full opt-in leads to higher individual accuracy than the control condition. We also found evidence that individual accuracy did not differ between the full and partial opt-in conditions.

Self-Direction is Beneficial to Crowd Performance

Table 3 shows crowd performance under the three performance metrics and summarizes the results of our analyses. In general, we found that self-direction is beneficial to crowd performance, with decisive evidence for the full opt-in crowd but mixed evidence for the partial opt-in crowd. We also found evidence that the full opt-in crowd and partial opt-in crowd do not differ in performance at the crowd level.

The full opt-in crowd tended to outperform the control crowd. Although there was evidence that crowd accuracy was equivalent for the full opt-in and control conditions in Experiment 2, the other more sensitive metrics provide strong evidence to the contrary. There was decisive evidence that the proportion better was greater than 50% in the full opt-in condition and very strong evidence that the proportion correct was higher than that of the control condition. This higher degree of evidence in favor of the alternative hypothesis should override the weaker evidence, based on an inefficient statistic, in favor of the null hypothesis.

The evidence comparing the partial opt-in crowd to the control crowd was mixed. Our three statistical tests comparing crowd performance yielded one result favoring the null hypothesis (Crowd Accuracy), one result favoring the alternative hypothesis (Proportion Better), and one result that does not favor either (Proportion Correct). These analyses taken together are sufficiently ambiguous to not adjust beliefs in either hypothesis.

Not shown in Table 3 is the comparison between full opt-in and partial opt-in. We found consistent evidence that the full opt-in and partial opt-in crowd performed equivalently. In particular, we found anecdotal evidence that crowd accuracy does not differ between the two experimental conditions (82.0 vs 76.0%, $t(99) = 1.713$, BF = 0.4527). When comparing the proportion better, we found substantial evidence that the rate does not differ from chance (53.33% of 75 BF = 0.1689). Similarly, we found substantial evidence that proportion correct is equal in the full and partial opt-in conditions (76.07 vs 73.82%, $t(99) =$ 0.9969, BF = 0.1792).

Question Choice Correlates with Difficulty Ratings

Participants in Experiment 2 chose questions for similar reasons as in Experiment 1. Difficulty ratings and choosing behavior were highly correlated in both the partial opt-in ($r = -0.887$, $N = 100$, BF $> 100,000$) and full opt-in ($r = -0.884$, $N = 100$, BF $> 100,000$) conditions. This corresponds to decisive evidence that participants tended to select questions that received low average difficulty ratings.

Discussion

Self-determination improves the crowd by contributing more knowledgeable members to queries requiring particular expertise. In Experiment 2, we found mixed evidence in our replication of Experiment 1a that compared partial opt-in crowds to control crowds on the easy question set. However, we found that allowing participants complete control is beneficial to aggregate performance when comparing the full opt-in and control conditions. Allowing for complete self-direction did not impact aggregate performance relative to partial self-direction. This null effect is noteworthy because there any many situations in which the test administrator may not have a sensible idea of how many questions each respondent should provide answers to—the results here suggest that this choice can be left up to the respondents with no negative consequence. Second, *forcing* an increase in the number of questions a respondent must provide answers to is almost certain to decrease accuracy. This can be easily envisioned by imagining a case in which respondents have to provide answers to all but one question. Accuracy could not be much different from the control condition, which was outperformed considerably in both experiments. Yet respondents who *choose* to answer additional questions detect that they are in a part of the quantity-accuracy trade-off function that is relatively flat—that is, they are increasing the quantity of their output without decreasing the accuracy. Having more responses is especially important in a pool of limited respondents or with questions of highly variable difficulty. Taken together, it would seem that allowing respondents to fully opt in, as they see fit, has several advantages and no obvious disadvantages.

Conclusions

Metacognitive choices about when to respond and what level of detail to provide typically enhance accuracy, a finding that implies that people have a good ability to assess what they do and do not know (Goldsmith and Koriat 2007; Koriat and Goldsmith 1996). Here, we explored whether individual metacognitive ability can be leveraged to enhance crowd wisdom. In two experiments, allowing individuals freedom as to which questions to answer led to a wiser crowd than constraining that freedom. The origin of this effect is in higher quality responses when people self-select items for which they have expertise.

The impressive benefits of allowing participants to opt in when aggregating likely depend on some factors relating to the task given to the participants. Under situations in which metacognitive monitoring is less accurate, the crowd will not benefit as greatly. And in cases where the items are constructed in such a way that accuracy and confidence are negatively related (Brewer and Sampaio 2012), then the self-determined crowd may actually be less wise than the control crowd. There are many domains in which metacognition has lesser benefits for the performance of individuals (e.g., perception-based tasks, Kunimoto et al. 2001), and so there would be no advantage for the crowd that allows self-direction. However, such cases are notably rare—in general, confidence is an exceptional predictor of accuracy in a wide range of circumstances (Wixted et al. 2015). For the ubiquitous domains where participants exhibit correct metacognitive judgments, crowd performance is likely to follow.

These findings imply that a design that aims to solve a set of problems via crowdsourcing would benefit from allowing users to select which tasks to solve. Design choices of this nature may also impact user experience and consequently influence how likely they are to use the platform. As such, even when the goal of a platform is to maximize performance over a set of questions, the degree of self-direction granted to users should be that which both benefits user experience and the quality of the resultant product. These findings support the design decisions behind online crowdsourcing platforms such as prediction markets and other crowd-sourced forecasting services where users have full control over the questions they answer. Such platforms contrasts to other forecasting approaches such as the “Delphi technique” (Hsu and Sandford 2007) where individuals of (putative) expertise in the domain of interest are assembled and decisions are reached through a combination of individual deliberation and consensus. Though the Delphi technique appears to be at least somewhat successful (e.g., Sniezek 1989 & Rowe and Wright, Rowe and Wright 1999), difficulties and costs with implementing such a technique are readily apparent. Here, we have shown that a simple manipulation imposed upon a less selected sample of respondents can serve the same purpose with little cost. Individuals are often the best judges of what they do and do not know—it only makes sense to leverage this metacognitive knowledge in search of wiser crowds.

Appendix

Experiment 1b judgments

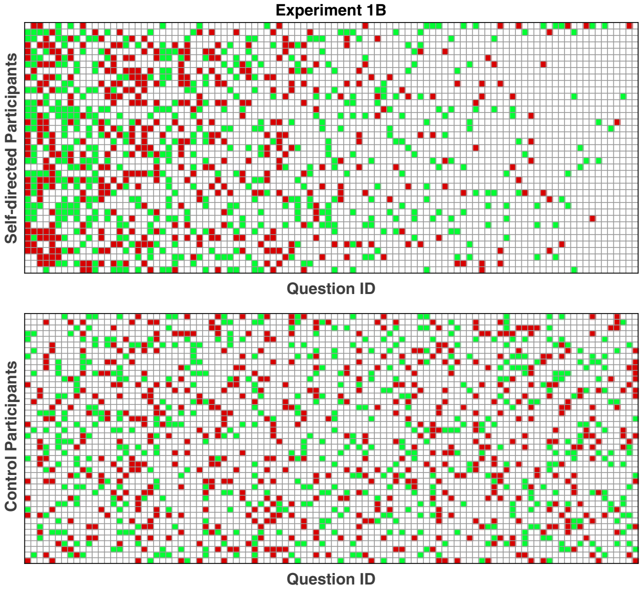

Fig. 4 Question responses for the self-directed and control participants in the hard condition with questions sorted by the number of participants who selected the question in the self-directed condition. Green squares represent correct responses, red squares represent incorrect responses, and white squares represent no response. Self-directed participants tend to cluster around the same collection of questions when compared to control participants

Experiment 2a judgments

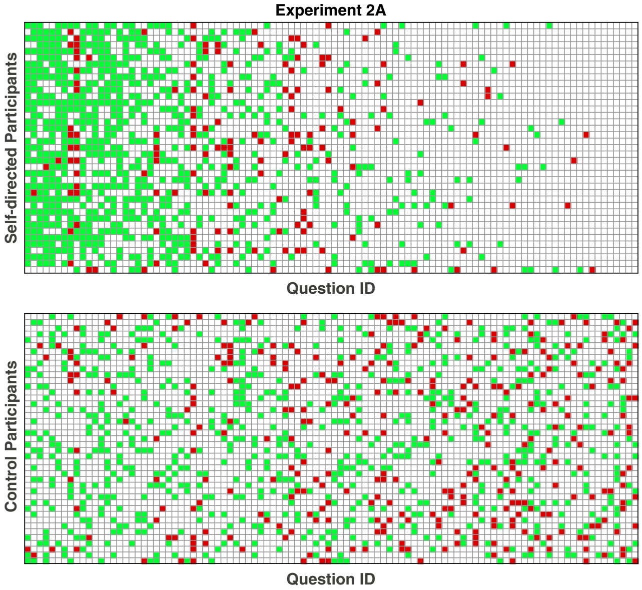

Fig. 5 Question responses from the partial opt-in and control conditions with questions sorted by the number of self-directed participants who selected the question. Green squares represent correct responses, red squares represent incorrect responses, and white squares represent no response. Self-directed participants tend to cluster around the same collection of questions when compared to control participants

References

Bedford, T., & Cooke, R. (2001). *Probabilistic risk analysis: foundations and methods*. Cambridge: Cambridge University Press.

Benjamin, A.S. (2008a). Memory is more than just remembering: strategic control of encoding, accessing memory, and making decisions. *Psychology of learning and motivation*, *48*, 175–223.

Benjamin, A.S., & Ross, B.H. (2008b). *The psychology of learning and motivation: skill and strategy in memory use* Vol. 48. New York: Academic Press.

Brewer, W.F., & Sampaio, C. (2012). The metamemory approach to confidence: a test using semantic memory. *Journal of Memory and Language*, *67*(1), 59–77.

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

Culpepper, S.A., & Balamuta, J.J. (2017). A hierarchical model for accuracy and choice on standardized tests. *Psychometrika*, *82*(3), 820–845.

Fiechter, J.L., Benjamin, A.S., Unsworth, N. (2016). *16 the metacognitive foundations of effective remembering*, (p. 307). Oxford: The Oxford Handbook of Metamemory.

Galton, F. (1907). Vox populi. *Nature*, *75*(7), 450–451.

Goldsmith, M., & Koriat, A. (2007). The strategic regulation of memory accuracy and informativeness. *Psychology of learning and motivation*, *48*, 1–60.

Hsu, C.-C., & Sandford, B.A. (2007). The delphi technique: making sense of consensus. *Practical assessment, research & evaluation*, *12*(10), 1–8.

Jarosz, A.F., & Wiley, J. (2014). What are the odds? a practical guide to computing and reporting bayes factors. *The Journal of Problem Solving*, *7*(1), 2.

Jeffreys, H. (1961). *Theory of Probability*. Oxford.

Kepecs, A., & Mainen, Z.F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1594), 1322–1337.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 609.

Kunimoto, C., Miller, J., Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*(3), 294–340.

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., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, *9*(3), 259.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A., Wagenmakers, E. (2018). Jasp (version 0.8.6). *Computer software. Retrieved from* https://jasp-stats.org.

Mannes, A.E., Soll, J.B., Larrick, R.P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S.E., Moore, D., Atanasov, P., Swift, S.A., Murray, T., Stone, E., Tetlock, P.E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. PMID: 24659192.

Merkle, E.C., Steyvers, M., Mellers, B., Tetlock, P.E. (2016). Item response models of probability judgments: application to a geopolitical forecasting tournament. *Decision*, *3*(1), 1.

Merkle, E.C., Steyvers, M., Mellers, B., Tetlock, P.E. (2017). A neglected dimension of good forecasting judgment: the questions we choose also matter. *International Journal of Forecasting*, *33*(4), 817–832.

Middlebrooks, P.G., & Sommer, M.A. (2011). Metacognition in monkeys during an oculomotor task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 325.

Olson, K.C., & Karvetsi, C.W. (2013). Improving expert judgment by coherence weighting. In *proceedings of 2013 IEEE International Conference on Intelligence and Security Informatics*.

Rocklin, T.R. (1994). Self-adapted testing. *Applied Measurement in Education*, *7*(1), 3–14.

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Rouder, J.N. (2018). Bayes factor calculator. Website. Retrieved from http://pcl.missouri.edu/bayesfactor. Accessed: 2018-04-23.

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

Schraw, G., Flowerday, T., Reisetter, M.F. (1998). The role of choice in reader engagement. *Journal of Educational Psychology*, *90*(4), 705.

Sniezek, J.A. (1989). An examination of group process in judgmental forecasting. *International Journal of Forecasting*, *5*(2), 171–178.

Steyvers, M., & Miller, B. (2015). Cognition and collective intelligence. In Bernstein, M., & Malone, T.W. (Eds.) *Handbook of Collective Intelligence* (pp. 119–138): MIT Press.

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

Surowiecki, J. (2004). *The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economics society and nations*. Brown: Little.

Tullis, J.G., & Benjamin, A.S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language*, *64*(2), 109–118.

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

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

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

Wixted, J.T., Mickes, L., Clark, S.E., Gronlund, S.D., Roediger, I. H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, *70*(6), 515.

Yi, S.K.M., Steyvers, M., Lee, M.D., Dry, M.J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science*, *36*(3), 452–470.