UNIVERSITY OF CALIFORNIA,
IRVINE

Metacognitively Wise Crowds

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Sciences

by

Stephen T. Bennett

Dissertation Committee:
Mark Steyvers, Chair
Joachim Vandekerckhove
Michael Lee

2020

Contents

Page

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank...

My parents Elizabeth and Chris, my brother David, and my sister Sarah, all of whom have all given me tremendous support.

My advisor Mark Steyvers, without whom I would have been perpetually lost.

Collaborators Aaron S. Benjamin and Percy K. Mistry who have helped make me a better writer and researcher. Some of the text of this dissertation is a reprint of the material as it appears in *Computational Brain and Behavior* and *Cognitive Science*. The co-authors listed in those publications directed and supervised research which forms the basis for the dissertation.

The journals of *Computational Brain and Behavior* and the society of *Cognitive Science* for allowing me to incorporate my work published with them into this dissertation.

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17071900005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

CURRICULUM VITAE

Stephen T. Bennett

EDUCATION

Doctor of Philosophy in Cognitive Science **2020**
University of California, Irvine

Master of Arts in Psychology **2016**
University of California, Irvine

Bachelor of Arts in Psychology **2014**
Lewis and Clark College *Portland, Oregon*
Bachelor of Arts in Mathematics **2014**
Lewis and Clark College *Portland, Oregon*

RESEARCH EXPERIENCE

Graduate Student Researcher **2016–2020**
University of California, Irvine

TEACHING EXPERIENCE

Teaching Assistant **2014–2016**
University of California, Irvine

REFEREED JOURNAL PUBLICATIONS

Making a wiser crowd: Benefits of individual metacognitive control on crowd performance **2018**
Computational Brain and Behavior

Leveraging metacognitive ability to improve crowd accuracy via impossible questions. **Under review**

REFEREED CONFERENCE PUBLICATIONS

A Bayesian model of knowledge and metacognitive control: Applications to opt-in tasks. **2016**
Cognitive Science

PREPRINTS

Estimating COVID-19 Antibody Seroprevalence in Santa Clara County, California. A re-analysis of Bendavid et al. **2020**
medRxiv

ABSTRACT OF THE DISSERTATION

Metacognitively Wise Crowds

By

Stephen T. Bennett

Doctor of Philosophy in Cognitive Sciences

University of California, Irvine, 2020

Mark Steyvers, Chair

Aggregates of many judgments tend to outperform each of the individual judgments that compose the aggregate, termed the Wisdom of Crowds effect. Metacognition has played an understudied role in the efficacy of these crowds and so in a series of experiments I explore how metacognition and self-direction can be used to improve crowd wisdom. I first demonstrate empirically that individuals can leverage their metacognitive abilities to improve the performance of crowds when they are allowed to opt-in to questions of their choosing. I develop a Bayesian framework wherein latent contextual knowledge describes how crowd members make opt-in decisions to elucidate the relationship between these cognitive and metacognitive processes. I then show that metacognitive ability can be estimated by asking questions with no correct response options and create metacognitively wise crowds which achieve more accurate responses despite incorporating fewer crowd members. I discuss my contributions to a geopolitical forecasting competition in which I developed models that combine human and algorithmic judgments to create highly accurate forecasts of the future. In this competition, I evaluated the effects of attenuating forecaster self-direction in an applied setting. These findings collectively demonstrate the importance of metacognition in forming accurate aggregate judgments and clarify the underlying metacognitive processes involved in self-direction.

Chapter 1

Overview

To help readers navigate this document, this introduction provides a brief overview of the topics and findings contained in each chapter.

In Chapter 2, I review the underappreciated role of metacognition in previous studies and applications of the wisdom of the crowd effect. I then establish the theoretical reasons that letting crowd members choose which questions to answer (i.e. opt-in) might either improve or detract from crowd performance. To resolve this uncertainty, I conduct an experiment and demonstrate that, letting crowd members opt-in improves the aggregate performance of crowds.

In Chapter 3, I develop a Bayesian cognitive model of the cognitive and metacognitive features of crowds that opt-in. This model leverages latent representations of knowledge to describe why individuals choose some questions over others and shows how self-direction can be expected to impact the performance of crowds.

In Chapter 4, I investigate how questions with no correct answers, termed impossible questions, can be used to measure metacognitive ability. I then demonstrate that in contexts where crowd members can opt-in, this metacognitive measure is a better predictor of a crowd member's contribution to the crowd than the direct measure of accuracy. As a result, highly accurate crowds can be created by relying on the most metacognitively capable individuals.

In Chapter 5, I detail my contributions to a Hybrid Forecasting Competition in which I developed human-algorithm hybrid models capable of outperforming either humans or algorithms alone. I show that incorporating algorithmic judgments in this way results in aggregate forecasts that are resilient to sparse human forecasts and therefore capable of scaling to large numbers of questions. Lastly, I evaluate the impact of an intervention in which crowd members were limited in which questions they could answer, identifying potential boundaries on the benefits of self-direction.

In Chapter 6, I discuss the general results of the research contained in the dissertation and promising future research directions that build upon these established findings.

While not directly related to the topics covered in this dissertation, Chapter 7 details a short re-analysis of COVID-19 infection rates using a simple Bayesian model that may be of interest to readers.

Chapter 2

Making a wiser crowd: Benefits of individual metacognitive control on crowd performance

This work was originally published in *Computational Brain and Behavior*.

Bennett, S. T., Benjamin, A. S., Mistry, P. K., & Steyvers, M. Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. *Computational Brain & Behavior*, 1(1), 90-99, 2018.

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.

2.1 Introduction

The earliest and most famous example demonstrating the wisdom of the crowd comes from a report by Francis Galton [33]. 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 [116], recitation of lists in an appropriate order [100], and in predicting events with as-yet unknown outcomes [56, 107, 71, 69].

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 [69] 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 [72]. 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 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 [103]. 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 [3], performance weighting [17], coherence and consistency weighting [78, 110] and consensus models [57]. For an overview of these methods, see [99]. 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 [36]. Control over selection of items for restudy [53] and over the allocation of study time [106] 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 [28, 7, 9].

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 [88, 94], and improve their performance by doing so. Psychometric models of this type of choice have demonstrated benefits in estimating subject characteristics [22]. Even non-human animals have the metacognitive ability to choose when to bet on their success in a particular trial [50, 74]. As noted above, the freedom to choose the grain size of reported memories substantially improves the accuracy of memories that are reported [36, 52]. 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.

2.2 Experiment 1

2.2.1 Method

Participants

166 participants were recruited through Amazon Mechanical Turk (AMT). Each participant was compensated $1 for the 30 minutes 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.

Table 2.1: Example questions. Correct answers are italicized.

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*

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 2.1. 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 5 questions to answer in that block. The participants in the control condition were randomly assigned 5 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 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.

2.2.2 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 [92, 42, 109], 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 [62] and a Bayes factor calculator available online [92, 91]. In both cases, we maintained the default priors that came with the software when performing computations.

In our notation, BF > 1 indicates support of the alternative hypothesis while BF < 1 indicates support of the null hypothesis. For instance, BF $= 5$ means the data support the alternative hypothesis by a factor of five. Similarly, BF $= 0.2$ corresponds to an equal amount of support of the null hypothesis. When discussing BFs, we use the language suggested by Jeffreys [44]. In order to improve readability, BFs larger than $100{,}000$ are reported as BF $> 100{,}000$.

Raw data

Figure 2.1 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 Figure A.1 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 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 2.2 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 2.3 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

Figure 2.1: 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.

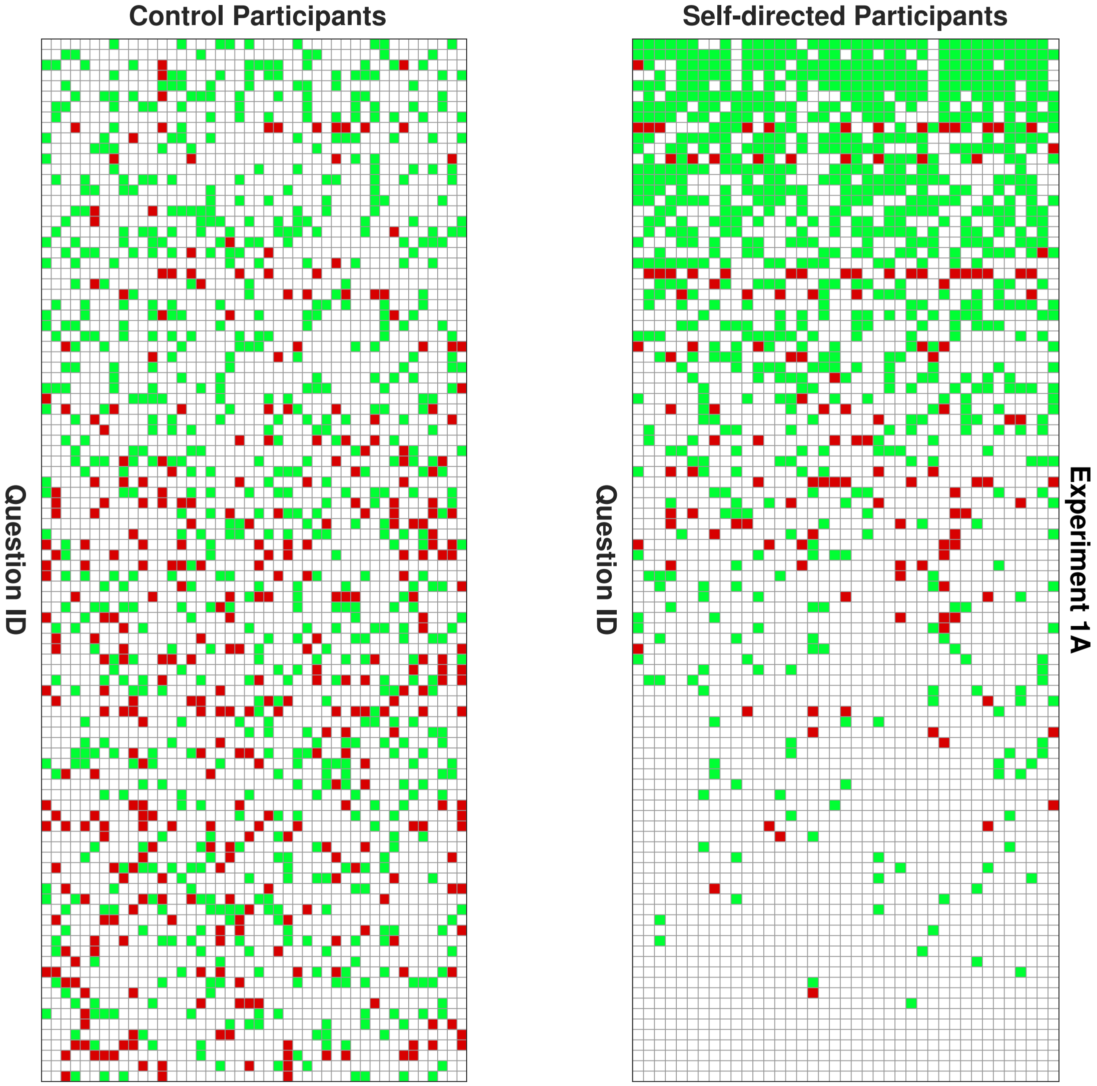


Table 2.2: Average individual performance across conditions. Each Bayes factor (BF) compares individual performance of the opt-in condition with the control condition within that experiment.

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

Table 2.3: Crowd performance across conditions and the performance metrics Crowd Accuracy, Proportion Correct and Proportion Better. Each Bayes factor (BF) compares performance of the opt-in condition with the control condition within that experiment.

	Crowd Accuracy				Proportion Correct					
	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

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 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

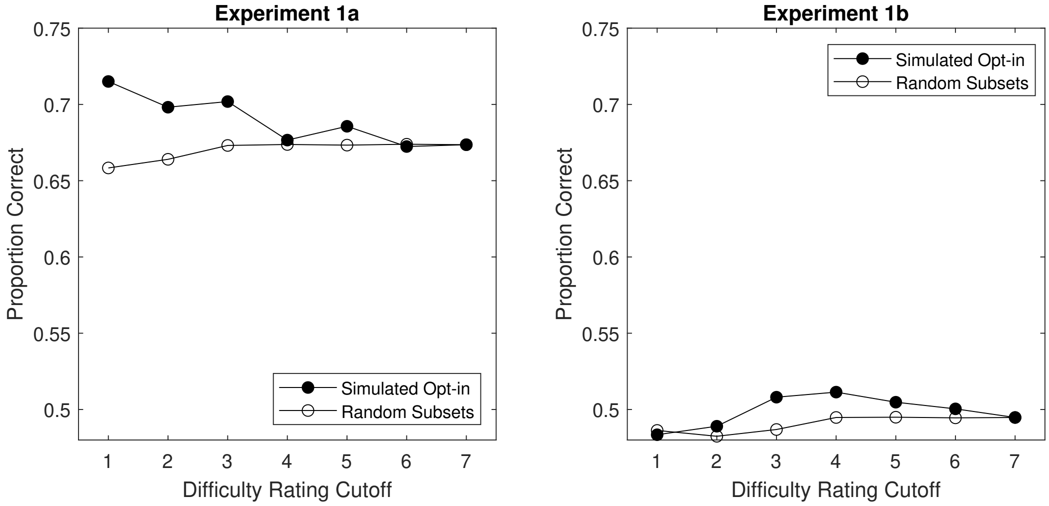
With the current data set, it is not possible to directly assess how participants would perform on questions that they did not choose. However, we can simulate an opt-in decision for participants in the control condition based on their difficulty ratings (participants rated all questions). Previously, it was found that crowds composed of confident responses led to higher accuracy than same-sized crowds composed at random [66].

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 2.2 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).

2.2.3 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

Figure 2.2: Crowd performance for subsets of judgments below a difficulty rating (simulated opt-in) and randomly chosen judgments (random subsets).



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.

2.3 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 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.

2.3.1 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 minutes 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.

2.3.2 Results

Responses from the full opt-in condition are shown in Figure 2.3 and those from the partial opt-in condition can be viewed in Figure A.2 in the Appendix. As in Experiment 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 2.2 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.

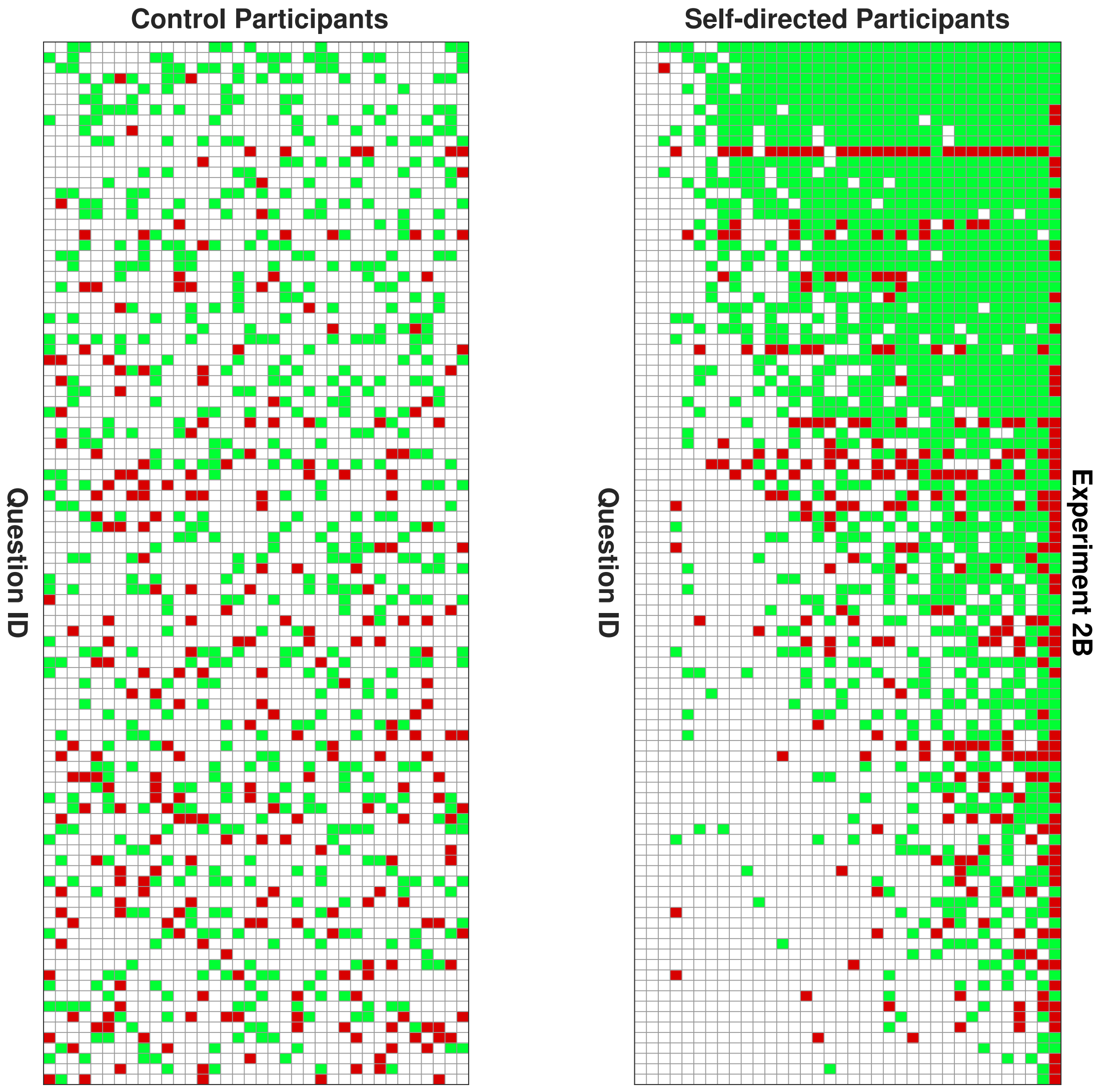
Self-direction is beneficial to crowd performance

Table 2.3 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.

Figure 2.3: 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.



Not shown in Table 2.3 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.

2.3.3 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.

2.4 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 [36, 52]. 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 [16], 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, [54]), 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 [114]. 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" [41] 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., [96] & [93]), 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.

Chapter 3

A Bayesian model of knowledge and metacognitive control: Applications to *opt-in* tasks

This work was originally published in *Proceedings for the 39th annual meeting of the Cognitive Science Society*.

Bennett, S. T., Benjamin, A., & Steyvers, M. A Bayesian model of knowledge and metacognitive control: Applications to *opt-in* tasks. *Proceedings for the 39th annual meeting of the Cognitive Science Society*, 2017.

In many ecologically situated cognitive tasks, participants engage in self-selection of the particular stimuli they choose to evaluate or test themselves on. This contrasts with a traditional experimental approach in which an experimenter has complete control over the participant's experience. Considering these two situations jointly provides an opportunity to understand why participants opt in to some stimuli or tasks but not to others. We present here a Bayesian model of cognitive and metacognitive processes that uses latent contextual knowledge to model how learners use knowledge to make opt-in decisions. We leverage the model to describe how performance on self-selected stimuli relates to performance on true experimental tasks that deny learners the opportunity for self-selection. We illustrate the utility of the approach with an application to a general-knowledge answering task.

3.1 Background

In traditional approaches to experimental psychology, an experimenter has unilateral control over which stimuli a participant experiences and the tasks that they complete. Yet in many real-world situations, such as providing ratings to videos on the Internet, the participant has some or even total control over the specific stimuli and tasks that they experience. The choice behavior underlying such self-selection is an important domain of study called *metacognition* [77], and the self-selection of activities or stimuli is specifically called *metacognitive control* [28, 29]. Some work on monitoring and control processes in memory tasks focused on confidence judgments as an indicator of self-selection questions [49, 52]. It is unclear precisely how this self-selection is generated, however. To better understand metacognitive control behavior, a model is needed that accounts for performance on the task of interest as well as the choice behavior that leads participants to select only some stimuli for exposure, evaluation, or testing.

The major difficulty of such an endeavor is that participants select tasks according to their interests and expertise, and so the data is missing in a nonrandom fashion (see [61], for a description of other missing data scenarios). Consequently, participants can only be compared and their performance fairly evaluated if a model is specified for the opt-in process. If a participant does not opt in to a particular question, then we simply do not see that participant's response to that question.

A starting point in explaining opt-in behavior is that participants have some meta-knowledge of what it is they already know, and use that knowledge effectively in service of ongoing learning. People provide higher assessments of their ability to answer inference questions in domains in which they have greater expertise [15], and learners often choose to engage more effective study techniques for material that is more difficult for them [8]. Memory reports are also considerably more accurate when respondents have the option of withholding answers that they are unsure of or of titrating the grain size of their answers to their perceived accuracy [36].

Self-regulated learning often has substantial benefits in educational contexts [73, 117, 13, 80]. Learners use meta-knowledge to allocate time, resources, and activities to an array of learning goals, and this application increases overall performance compared to learners who have their learning activities dictated by an instructor [113, 29].

The benefits of self-control extend beyond these constrained tasks, however. In causal reasoning experiments, participants can more quickly understand the causal structure of a network if they intervene in the learning process and design their own “experiments” [101, 97, 55]. Human strategy selection can be explained in terms of rational metareasoning, wherein humans flexibly choose strategies in accordance with their environment [59, 60].

The core claim across each of these examples is that self-selection within a task aimed at measuring performance is driven by metacognitive knowledge, which leads to a higher rate of success, expertise, or interest for the selected items. This process makes it difficult to evaluate the stimuli and the participants in an unbiased way. One test-taker may, for example, outperform another not because they have greater knowledge but rather because they make more a judicious selection of problems.

The aim of this project is to develop a cognitive model of the metacognitive aspect of item selection. In doing so, it also provides a framework to relate performance on self-selected materials with performance on an unconstrained set of items or stimuli. Here we apply this model to data collected from participants answering general knowledge questions, but the model is considerably more general: the same principles could apply in other metacognitive control tasks, such as study time allocation or selection of items for restudy. We are aware of one current model of metacognitive

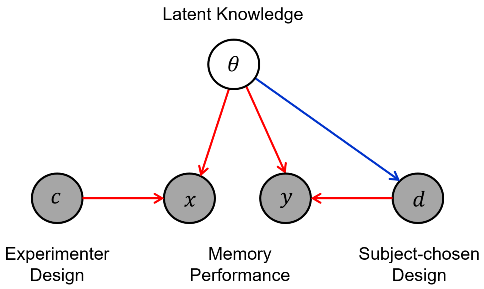


Figure 3.1: Outline of our modeling approach. Latent knowledge and design both explain the performance on the task. In the case of a subject-chosen design, latent knowledge also explains the design.

control, which takes as a given the state of the world, which then causes the observed behaviors [30]. We take a different approach, which starts with the latent knowledge that the participant is coming to the experiment with and uses that in both the selection process behind opting-in and the observed responses to questions, as illustrated in Figure 1. Performance on the task is explained by both the design—that is, the particular experience of the participant in the task—and the latent knowledge of the participant. In the case of a participant who can opt in to certain questions but avoid others, the design is also partially informed by the latent knowledge. We are interested in estimating the latent knowledge of each participant and evaluating how it relates both to performance on the task and to opt-in behavior. In order to infer latent knowledge from the observed data, we apply Bayes' rule in the equation below, where θ is the latent knowledge, c is the experimenter design, d is the subject-chosen design, x are the performance data from a true experimental design (where subjects respond to all or to a random subject of probes), and y are the performance data from a subject-chosen design (in which subjects choose which probes to respond to):

$$p(\theta|c,d,x,y) \propto p(x|\theta,c)p(y|\theta,d)p(d|\theta)p(\theta) \tag{3.1}$$

In a traditional cognitive model, the important part of the model is the specification of $p(x|\theta,c)$ and $p(y|\theta,d)$, termed the likelihood functions. These functions directly explain the empirical effect of interest by relating latent knowledge to performance on the task given the experimental design. The novel part of the model relates to the specification of the metacognitive control process $p(d|\theta)$, which explains how the participant self-designs on the basis of their latent knowledge. If we would ignore this model component, we would likely, and incorrectly, conclude that participants who self-designed were more knowledgeable than participants subject to the experimenter's design because they outperformed their experimenter-designed counterparts. Such an error could be catastrophic if we were trying to compare across individuals or across tests. Because subjects are randomly assigned to conditions, it is highly unlikely that they differ widely. The process by which the participants who self-designed outperformed those who could not lies in the opportunity to self-design. Here we see the importance of jointly modeling the selection process and the task at hand in order to understand the interplay between latent knowledge, opt-in behavior, and performance.

Since this is a task in which many participants give judgments to many questions, we also expect to find that averaging across participants leads to higher accuracy–an effect termed the *wisdom of the crowd* [103, 100]. Here we have the opportunity to evaluate whether the opportunity to opt in to a self-selected portion of the questions will enhance or attenuate such benefits associated with averaging. Certainly, many participants will gravitate towards the same questions when they can opt in, which would potentially decrease the benefits of averaging across a crowd by virtue of reducing input to the more difficult questions. However, based on what is known about metacognition, we expect that participants will opt in to questions for which they have relevant knowledge, which could lead to a more informed set of responses to average with the remaining crowd. Crowd behavior provides an additional benchmark against which we can evaluate the performance of the metacognitive model.

Table 3.1: Example questions.

Difficulty	Example
Hard	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
Easy	Greenhouse effect refers to: (a) gases in the atmosphere that trap heat, or (b) impact to the Earth's ozone layer

3.2 Experiment

Stimuli The question set consisted of 100 general-knowledge binary choice questions. The questions were drawn from 12 topics: World Facts, World History, Sports, Earth Sciences, Physical Sciences, Life Sciences, Psychology, Space & Universe, Math & Logic, Climate Change, Physical Geography, and Vocabulary. The question set was created by collecting from multiple sources. Two example questions are shown in Table 3.1. Based on the empirically observed accuracy levels, the first is difficult and the second is easy.

Participants A total of 83 participants were recruited through Amazon Mechanical Turk (AMT). Each participant was compensated $1 for the 30 minutes the experiment was expected to take, and assigned to one condition.

Design Participants could view the survey description on AMT. If they selected the survey they were redirected to another website. They were first directed to a study information sheet which provided details of the survey and compensation. If they agreed to continue, they were instructed to answer some demographic questions. Participants were randomly assigned to either a *random* condition ($N = 44$) or a *self-selection* condition ($N = 39$), determining the subject's role in selecting which questions to answer. Participants were not aware of the existence of other conditions. Each participant saw the questions in 5 blocks of 20 questions each. In each block, they were instructed to rate the difficulty of each question and then, if they were assigned to the opt-in condition, instructed to choose 5 of those 20 questions to answer. The participants in the random assignment condition were randomly assigned 5 questions from that block to answer. After rating the difficulty of all 100 questions and answering 25 of them, participants were thanked for their time and given instructions on how to receive payment.

3.3 Model

The model utilizes an IRT model to generate subjective latent knowledge (the belief of a participant that she can answer a question), which informs all aspects of participants' responses including the observed accuracy and difficulty ratings, as well as the metacognitive process of question selection. We describe participants as opting-in to questions for which they believe they have knowledge, answering with accuracy dependent on whether or not they believe they have knowledge, and giving lower difficulty ratings when they believe they have knowledge.

We use an IRT model to generate the subjective latent knowledge, $\delta_{i,j}$, for each participant i (across both the opt-in and random condition) and question j,

$$\delta_{i,j} \sim \text{Bernoulli}(\text{logit}^{-1}(\theta_i + \eta_j)) \tag{3.2}$$

where θ_i is the self-perceived skill of participant i, η_j is the perceived familiarity of question j, and $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$. This latent knowledge is represented as a 0 or 1, indicating whether or not that participant believes that she has knowledge for that question. We place a Normal prior on the self-perceived skill, $\theta_i \sim \text{Normal}(0, \sigma)$, such that participants are expected to have the same skill (on average) for both the self-selection and random conditions.

For the self-selection condition, we assume that participants have a preference to select questions for which they believe they have knowledge. Let c represent the observed question selections with $c_{i,j} = 1$ if question j was selected by participant i. For each participant and question block, we model question selection in the opt-in condition by a sampling process:

$$c_i \sim \text{SampleWR}((\delta_{i,1}+\kappa,...,\delta_{i,K}+\kappa),M) \tag{3.3}$$

where K is the total number of questions available for selection in each block (K=20 in our experiment), M is the number of questions that need to be selected (M=5 in the experiment), SampleWR(δ,M) represents a sampling without replacement distribution where M items are sampled with probability proportional to δ, and κ is a fixed parameter that controls the randomness in the selection process. Higher κ values make it more likely that questions are selected for which the participant has no subjective knowledge. For participants in the random condition, we assume that the questions are randomly sampled by a process that is under control of the experimenter (where M out of K questions are randomly allocated).

Let $x_{i,j}$ represent the observed accuracy for participant i on question j. We do not assume a fixed relationship between belief of knowledge and accuracy. For each question, we introduce guessing rate parameters ρ_j and λ_j that control the probability of correct responding if the participant does or does not have subjective knowledge about a question:

$$x_{i,j} \sim \text{Bernoulli}(\delta_{i,j}\rho_j + (1-\delta_{i,j})\lambda_j) \tag{3.4}$$

For example, with $\rho = 0.8$ and $\lambda = 0.4$, the probability of a correct response is 0.8 if a participant has subjective knowledge, but 0.4 if the participant does not. The guessing parameters are given Beta priors, $\rho_j \sim \text{Beta}(\alpha,\beta)$, $\lambda_j \sim \text{Beta}(\alpha,\beta)$ where α and β are hyperparameters that control the variability in guessing rates across questions.

To model the difficulty ratings, we use an ordered logit model [112]. We assume that subjective latent knowledge informs the perceived difficulty of questions. Questions for which the participant believes they have knowledge are perceived as easier. Let $\phi_{i,j}$ represent the perceived difficulty for participant i on question j. We determine the perceived difficulty by:

$$\phi_{i,j} = -\delta_{i,j} - \eta_j \xi + \omega_i - \beta_j + \sigma_{i,j} \tag{3.5}$$

where β_j and ω_i capture participant and item level effects (e.g. some participants might find all items easy, some items might be judged as easy) independent of subjective knowledge. In addition, we also allow the perceived familiarity of a question η_j to affect the perceived difficulty weighted by a fixed scaling parameter ξ. Finally, $\sigma_{i,j}$ represent small perturbations centered around 0 to explain the random variability in difficulty ratings unrelated to any of the previous factors mentioned. These perceived difficulties feed into the ordered logit model to generate the difficulty ratings $r_{i,j}$,

$$r_{i,j} \sim \text{OrderedLogit}(\phi_{i,j}, \tau_i) \tag{3.6}$$

where τ_i is the set of criteria cutoffs for participant i.

We used JAGS to perform parameter inference. All parameters were inferred jointly from the opt-in and random condition. All model predictions were derived from posterior predictives where we simulate new participants from the distribution and assess how they self-select from a new set of questions.

3.4 Results

We examine several empirical effects within the data and observe that the model captures the appropriate trend in most cases.

Item selection and latent knowledge. The model captures the expected relationship between opting-in behavior and knowledge (see Figure 3.2). Participants were more likely to select questions for which they had pre-existing knowledge. Each question was randomly assigned to at least four participants in the random assignment condition. However, in the opt-in condition, there were seven questions that no participant chose to answer. Question selection strongly corresponded with the inferred latent knowledge ($\delta_{i,j}$) for the participant-question pair, with participants choosing questions for which they had latent knowledge. Across conditions, latent knowledge is distributed in a similar manner: most participants have knowledge for popular questions, few participants have knowledge for unpopular questions, and some participants are more knowledgeable than others. However, the model has substantially more certainty about the localization of this knowledge in the opt-in condition compared to the random condition because it can leverage the opt-in behavior. In Figure 3.2, this certainty is expressed as black or white squares, while uncertainty is represented in gray. We see the uncertainty about which participants have knowledge for which question as a "blurring" of the latent knowledge space.

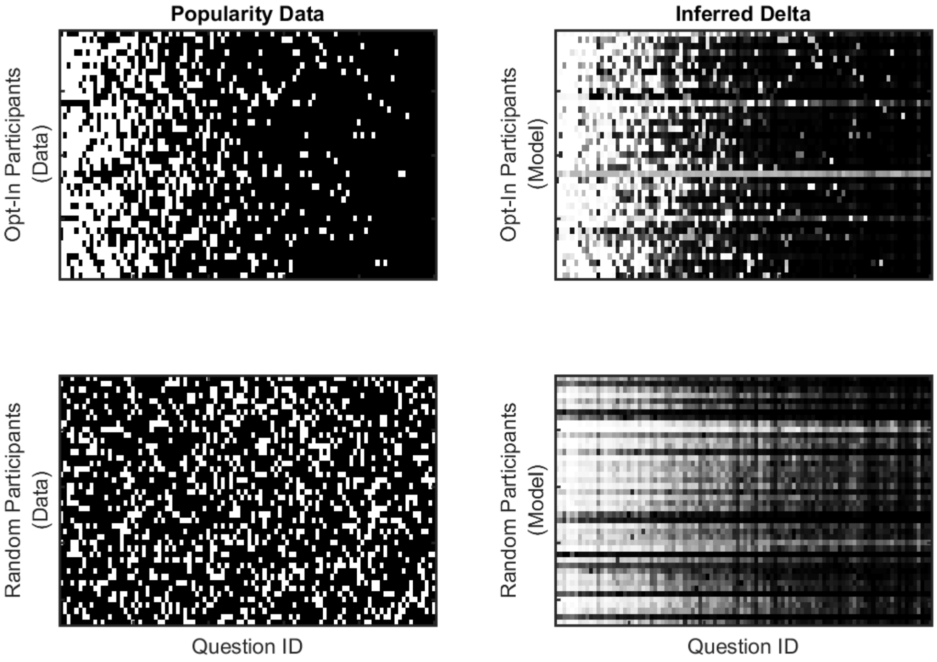


Figure 3.2: Latent knowledge is similar between conditions and corresponds to opting-in behavior. Plotted are the opt-in behaviors and average $\delta_{i,j}$ values across conditions, all sorted by the popularity of the question in the opt-in condition. White corresponds to questions that the participant opted in to or the inferred presence of knowledge.

Effect of opting-in on participant performance. Average performance across questions was higher in the self-selection condition (86.05%) than in the random condition (67.27%). We computed a Bayes Factor (BF) given a binomial distribution with a shared or different rate of correct responding and find a Log_{10} BF of 21.12 in favor of a higher rate of correct responding in the opt-in condition. This corresponds to decisive evidence that average accuracy is higher in the opt-in condition than the random assignment condition. This occurs even when taking into account the fact that people tend to opt in to easier questions. In order to perform this analysis, we took the product of the evidence that performance is higher in the opt-in condition than the random assignment condition for each question and find a Log_{10} BF of 9.02. So, even when comparing on an item-by-item basis, opting-in provides an advantage.

Effect of opt-in on model performance. For the model, the average accuracy for posterior predictive samples in the self-selection condition (mean = 79.03%) is also significantly higher than in the random condition (mean = 67.07%), both across all questions (99.86 % of samples) and even within questions (68.93 % of sample-question pairs). We observe this benefit in accuracy despite the average inferred ability of individual subjects (θ_i) being equivalent across conditions: $\overline{\theta}_i = 0.00$, SD = 0.99 in the opt-in condition versus $\overline{\theta}_i = -0.09$, SD = 2.04 in the random assignment condition. This means that the benefit to accuracy that the model predicts is due to downstream consequences of the metacognitive selection process and not an (inaccurate) inference that participants in one condition were more skillful than in the other.

Difficulty Ratings. Participants tended to give lower average difficulty ratings to questions that they opted in to (Log_{10} BF = 91.89) and higher average difficulty ratings to questions that they did not opt in to (Log_{10} BF = 64.09), relative to the random condition. The model captures, but understates, this trend (see Figure 3.3).

Wisdom of the crowd. The left panel of Figure 4 shows the relationship between crowd size and crowd accuracy for the two conditions in the experiment, as well as a hybrid condition in which the two groups are combined. The right side of the Figure shows that the model captures this effect

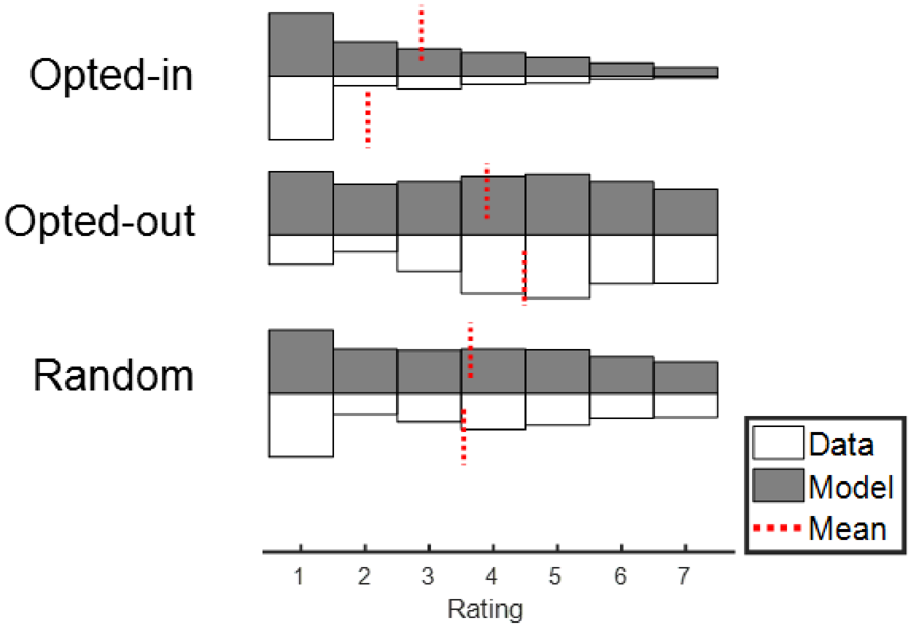


Figure 3.3: Distribution of difficulty ratings for participants and model for questions that were selected or not selected in the opt-in condition and the random condition. Lower ratings indicate lower perceived difficulty.

qualitatively. Crowd responses were determined by taking the most common response across the participants in the crowd. Since seven questions went unanswered in the opt-in condition, we had to consider how unanswered questions impacted crowd performance. To treat the self-selection condition maximally conservatively, we graded any question that went unanswered as incorrect for that crowd. Even with this penalty, the crowd composed of the participants from the self-selection condition (79%) outperformed the crowd of subjects from the random condition (73%).

We also considered the impact of crowd size on performance. To do this, we evaluated the average performance of crowds composed of random samples of participants from a condition and varied the number of participants drawn to form the sample. We plot average crowd performance as a function of the total number of judgments, where a judgment is a person's response to a question. The hybrid condition provides a means of improving upon both conditions. To create a hybrid crowd, we first sampled participants that answered the question from the opt-in condition. If a question had no responses, we added the answer from one participant in the random condition in order to guarantee that all questions received at least one answer. This hybrid crowd has high performance across all questions. The model captures the general trends in the data in that larger crowds result in higher crowd accuracy, opt-in crowds outperform random-assignment crowds, and the hybrid crowds perform well across all questions.

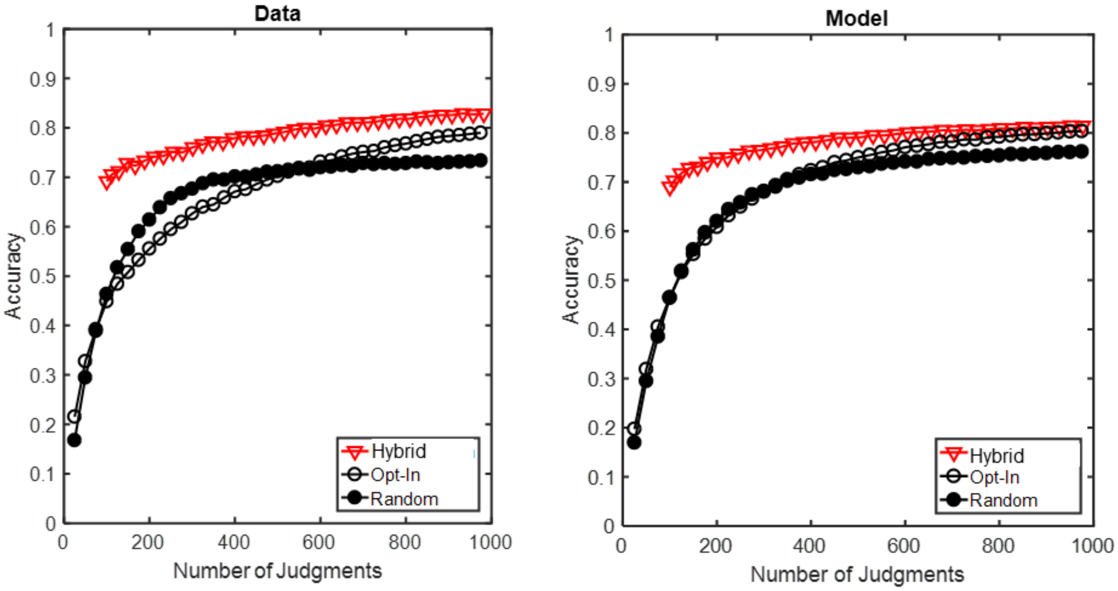


Figure 3.4: Crowd performance when varying the number of participants (measured by the total number of judgments)

Additional simulations. Given our model, we investigated which circumstances would likely lead to changes in the relative performance of the self-selection and random conditions in terms of both average overall accuracy and crowd performance. We varied the heterogeneity of perceived question difficulty (η_j) and latent ability (θ_i). We did this by simulating experiments in which we varied the underlying hyper-parameter corresponding to the variability of θ_i and η_j by factors of 0.25, 1, and 4 while keeping other parameters constant (see Figure 3.5). We find that increasing the heterogeneity of perceived question difficulty increases self-selection accuracy overall, but decreases it at the crowd level since participants tend to avoid answering the same difficult questions. Heterogeneity in question difficulty does not have an appreciable impact on performance in the random condition. In both conditions, higher heterogeneity of participant skill leads to higher crowd performance and gives resilience to heterogeneously difficult questions in the opt-in condition. However, it detracts from overall accuracy in the self-selection condition.

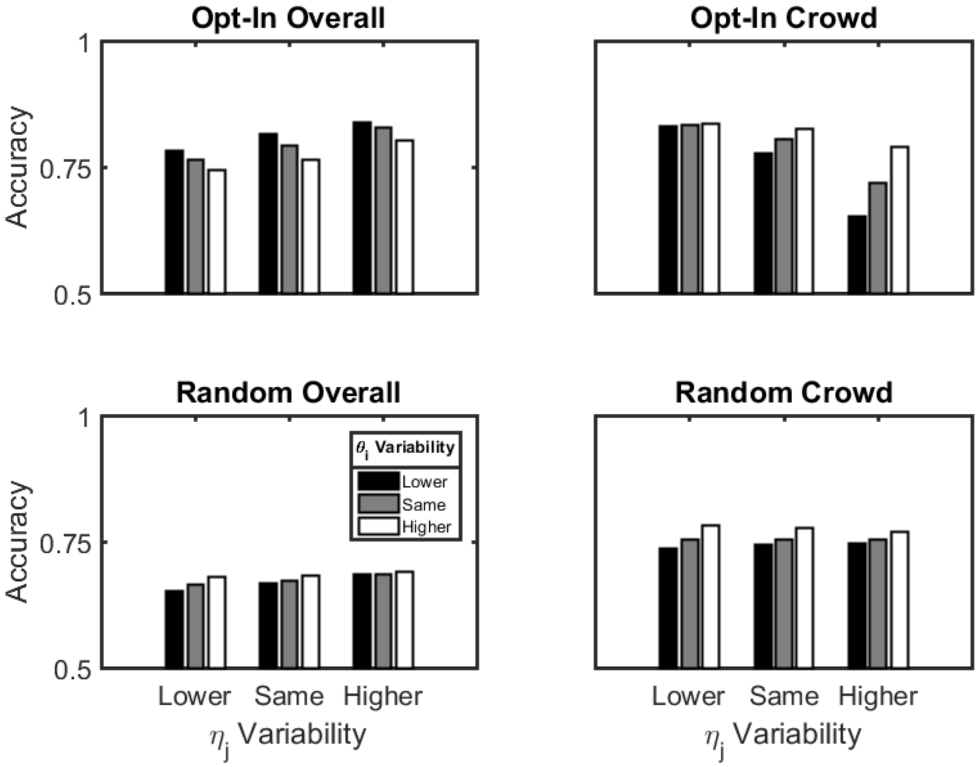


Figure 3.5: Simulated performance depending on variability of question difficulty (η_j) and participant skill (θ_i).

3.5 Conclusions

A comprehensive model of cognition must make allowance for the fact that cognitive behavior is driven by motivations. We choose what we attend to and attempt to encode, and what we attempt to remember. Metacognitive behavior is at the heart of most learning outside the laboratory, and a fair amount within it as well [6]. The joint modeling of metacognitive behavior–like self-selection of items–along with cognitive performance has the potential to address a wider and more representative range of real-world learning and testing behaviors, and can serve as the basis for drawing comparisons across individuals or tests that would otherwise be hopelessly confounded. Additionally, the model could be extended to explain various incentives given to the participant, which would impact how latent knowledge interacts with the task to generate opt-in behaviors. The model presented here provides a starting point for such an enterprise. It leads to a relatively good description of performance across a variety of metrics. A single latent knowledge state for each participant-question pair permits an explicit representation of the metacognitive process that governs the relationship between opt-in, accuracy, and difficulty behaviors. The model is successful in describing the nonrandom missing nature of the data that we observed by relying on principled psychological theories about why someone might choose one question over another.

An additional lesson of the current research can be seen in the crowd data. Opting in is generally beneficial to crowd accuracy in both the observed data and our model. This result indicates that the metacognitive skill of the individuals in self-selection can be leveraged in order to create a smarter crowd. This effect is sufficiently robust that it appears to outweigh the cost associated with small crowd sizes for some questions or no volunteered responses at all for a small number of questions. Such a result is particularly important when considering the widespread availability of datasets in which responses are self-selected.

Chapter 4

Leveraging metacognitive ability to improve crowd accuracy via impossible questions.

This work is under review for publication.

Bennett, S. T. & Steyvers, M. Leveraging metacognitive ability to improve crowd accuracy via impossible questions. *Manuscript submitted for publication*.

The aggregate of judgments across individuals can be quite accurate, especially when individuals with expert judgment can be identified. A number of procedures have been developed to identify expert judgments using historical performance or questionnaire data. Here we identify expertise with the ability to decline answering impossible questions. These questions which have no correct answers serve as a metacognitive measure of a participant's ability to recognize when they lack knowledge, which is especially valuable in contexts where individuals choose which questions they answer and selectively contribute to the crowd's answer. We find that an individual's propensity to skip impossible questions is related to their expertise and leverage these questions to form highly accurate crowds, outperforming other methods of identifying experts that rely on historical accuracy.

4.1 Background

The aggregate of answers across individuals tends to be more accurate than most of the individual answers. Crowds can accurately predict the outcome of future geopolitical events via opinion pooling [107, 1, 4] or prediction markets [115, 98]. Crowds are used to generate accurate labels for images [111], EEG components [83], music [20], and medical segmentations [38]. While geopolitical forecasting and labeling are common applications, crowds are effective for a surprising breadth of tasks, including solving combinatorial problems [116], predicting the outcome of sporting events [39, 81], identifying authorship from handwriting [67], and visual search [47]. Despite the impressive history of crowds, aggregating across a large and diverse group does not guarantee accuracy [23, 19, 37].

One way to maximize the accuracy of a crowd is to identify experts within the crowd who provide consistently accurate responses. Many methods have been developed to identify and weigh experts, including those that rely on absolute accuracy [66], contribution weighted scores [21], or genetic algorithms [40]. If historical accuracy is difficult or costly to obtain, questionnaires such as the cognitive reflection task [32] allow the experimenter to select for more deliberative reasoners, which can be used to improve crowd accuracy [68, 26].

Another approach to improve crowd accuracy is to permit crowd members to select which questions to answer, or *opt-in* [10]. To the extent that crowd members have the metacognitive ability to assess their own competency for questions, the resultant crowd is highly accurate. Therefore, when crowd members are allowed to selectively contribute to different problems, their observed performance will be a combination of their domain expertise and metacognitive ability.

In many crowd contexts, metacognition plays a key role in crowd performance. In a geopolitical forecasting context, the most accurate crowd members selected a much broader range of questions than their peers [72], indicating that metacognitive ability is related to forecasting ability. Additionally, respondents' expectations about the distribution of other judgments can be used to identify relative experts within a crowd and make accurate predictions even when most responses are wrong [84, 85]. While this is not metacognition in the traditional sense of reasoning about one's own reasoning, this meta-knowledge of *others'* reasoning is a useful predictor of accuracy. While metacognition and domain skill are psychologically distinct entities, they are highly correlated [46]. Psychometric methods of analyzing metacognitive ability can reliably distinguish these traits, but require involved questionnaires [51]. *Meta-d'* is another method that can differentiate between competency and metacognitive ability using a signal detection framework [64], but requires a large number of responses to questions with known ground truth. Therefore, there is a need for simple methods that can assess metacognitive ability in cases where the ground truth is unavailable. Impossible questions may be able to fill this gap.

Impossible questions are questions for which no correct answer can be given. One form of these questions asks for details about a non-existent subject, such as the symptoms of *Seradot's disease*. To our knowledge, no such disease exists, so all statements about the symptoms of the disease are incorrect. These questions can serve as a pure metacognitive measure; no participant, no matter how knowledgeable, can do anything other than profess their ignorance without asserting a falsehood. Importantly, they are also questions for which there is no uncertainty on the part of the experimenter that the participant has any knowledge relating to the answer. The experimenter knows that the participant knows nothing about the subject. As a result, any deviation from maximum uncertainty is a known metacognitive error on the part of the participant.

A number of previous studies have utilized impossible questions. [12] termed them unsolvable items and focused on the impact of scoring rules on response strategies. Other studies have focused on overclaiming, where individuals assert familiarity with fictitious terminology in domains such as finance and biology [2]. Overclaiming can be used to readily identify when people overstate their own knowledge and abilities [82, 11, 24].

To our knowledge, no study has examined the use of impossible questions (or overclaiming) in a crowd setting. In crowds that opt-in, how do participants' responses to impossible questions relate to their contribution to the crowd? We conduct two experiments and re-analyze an existing data set by [12] to examine whether impossible questions can measure metacognitive ability and be used to improve crowd accuracy.

4.2 Experiment 1

4.2.1 Method

Participants

35 participants were recruited through Amazon Mechanical Turk (AMT). Using MTurk worker requirements, we restricted the participant pool to individuals living in the United States who had a 98% or higher HIT approval rating on at least 1000 HITs and had not participated in any of our previous studies that used overlapping questions. Each participant was compensated $5 for the 30 minutes the experiment was expected to take.

Stimuli

Stimuli consisted of 94 general knowledge binary-choice questions with a known ground truth. 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. Based on a previous experiment, this set of questions resulted in an average accuracy of 76% [10]. In addition to the general knowledge questions, participants were asked 6 binary choice impossible questions interspersed at random with the general knowledge questions. These were questions based on made-up concepts and so had no correct answers. Examples of both types of questions are shown in Table 4.1.

Table 4.1: Example general knowledge and impossible questions. Correct answers are italicized. Impossible questions have no correct answers.

Type	Example
General Knowledge	Greenhouse effect refers to: (a) *gases in the atmosphere that trap heat*, or (b) impact to the Earth's ozone layer?
	House flies have an average life span of less than 2 days. (a) True, or (b) *False*?
Impossible	What is the most prominent symptom of Seradot's disease? (a) A fever, or (b) A rash?
	Resistance Configuration Theory is a psychological theory that explains: (a) How people avoid blame and why they do not recognize when something is their fault, or (b) Why certain people do not try new experiences?

The full list of questions is available on the Open Science Framework (anonymized for review): `https://osf.io/8r3t9/?view_only=c5a9694d4900431ab00566e124a10b1d`.

Participants could opt-in or opt-out of each question. They opted-in by selecting either of the answers or skipped the question by selecting "opt-out." When a participant opted-out of a question, they did not answer it and that question had no impact on their displayed accuracy. For the impossible questions, all answers to the question were incorrect and so the only way to avoid an incorrect response was to opt-out. Whenever a participant opted in to a question, they were also required to report how confident they were in their answer, from 50% to 100% (participants who opted out were able to give a confidence rating, but were not not required to do so).

Design and Procedure

Participants could view the survey description on AMT. If they selected the survey they were first directed to a study information sheet that provided details of the survey and compensation. If they agreed to continue, they were shown an example question with instructions on how to navigate the experiment. After receiving instruction, participants answered or opted out of each of the 100 questions (94 general knowledge questions and 6 impossible) and then gave general confidence ratings for each of the 12 categories of questions.

Scoring Crowd Performance

We used two methods for evaluating crowd accuracy taken from [10]: Proportion Correct and Proportion Better. *Proportion correct* measures the accuracy of responses within each question and is the average of these accuracies. *Proportion better* compares two crowds; it is the proportion of questions for which one crowd has higher accuracy than the other.

Bayes Factors

For all 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 [48, 92, 42, 109], 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 [43] and a Bayes factor calculator available online [92, 90]. In both cases, we maintained the default priors that came with the software when analyzing data.

To interpret these Bayes Factors, we denote support for the alternative hypothesis with BF > 1 while BF < 1 indicates support of the null hypothesis. For instance, BF = 5 means the data support the alternative hypothesis by a factor of five. BF = 0.2 corresponds to an equal amount of support for the null hypothesis. When interpreting BFs, we use the language suggested by Jeffreys [45]. In order to improve readability, BFs larger than 100,000 are reported as BF $>$ 100,000.

4.2.2 Results

Data from all experiments reported in this article are publicly available on the Open Science Framework: `https://osf.io/8r3t9/?view_only=c5a9694d4900431ab00566e124a10b1d`.

Data Overview

The pattern of answers across participants and questions is shown in Figure 4.1. The figure shows the heterogeneous response patterns across participants and questions. Participants had an overall accuracy of 73.9% on the 94 general knowledge questions and opted in to those questions 91.6% of the time. Average confidence was higher when participants opted in than when they opted out[1] (85.6% vs 55.7%, t = 19.5, BF $>$ 100,000).

Impossible Questions

Participants opted out of the impossible questions much more frequently than they opted out of the general knowledge questions (8.4% vs 43.3%, BF $>$ 100,000 via Bayesian Contingency Table).

We term the number of impossible questions that the participant chose not to answer the Impossible Question Criterion (*IQC*). Higher IQC indicates a higher level of metacognitive ability. Participants with higher IQC were more accurate overall (r = 0.48, BF = 11.5) and more likely to opt out of the 94 general knowledge questions (r = 0.66, BF $>$ 880.0). Figure 4.2 shows these relationships.

[1] Confidence ratings were only required when the participant opted in and were optional when participants opted out. Collectively, participants gave confidence ratings for 118 of the 366 opt-out responses.

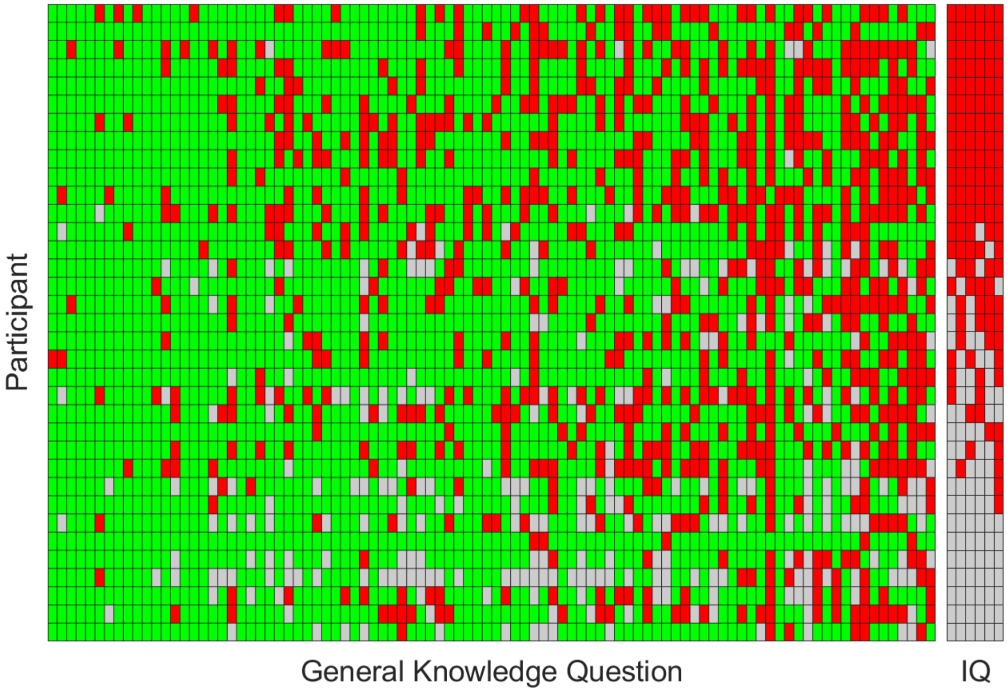


Figure 4.1: Participant responses to each question in Experiment 1. Green and red colors indicate a correct and incorrect response respectively. Grey colors indicate that the participant opted-out. Questions are sorted from left to right by increasing question difficulty as established by a previous experiment. Impossible questions are labeled "IQ". Participants are sorted by the number of impossible questions answered.

Selecting participants to create more accurate crowds

How can impossible questions be leveraged to form more accurate crowds? Since the individuals who correctly opt-out of the impossible questions have higher accuracy and less bias, we can use the impossible questions as a filter, allowing only those participants who opt-out of a sufficient number of impossible questions into the crowd. Using impossible questions as a filter in this way improves crowd performance relative to an unfiltered crowd (see Table 4.2).

However, there are other methods that could be used to select for high quality crowd members. Many online behavioral experiments use attention checks to filter out respondents with low quality responses. While we did not include any questions specifically designed to catch low-attention work, our set of questions covered a wide range of difficulties. The 6 easiest questions had an

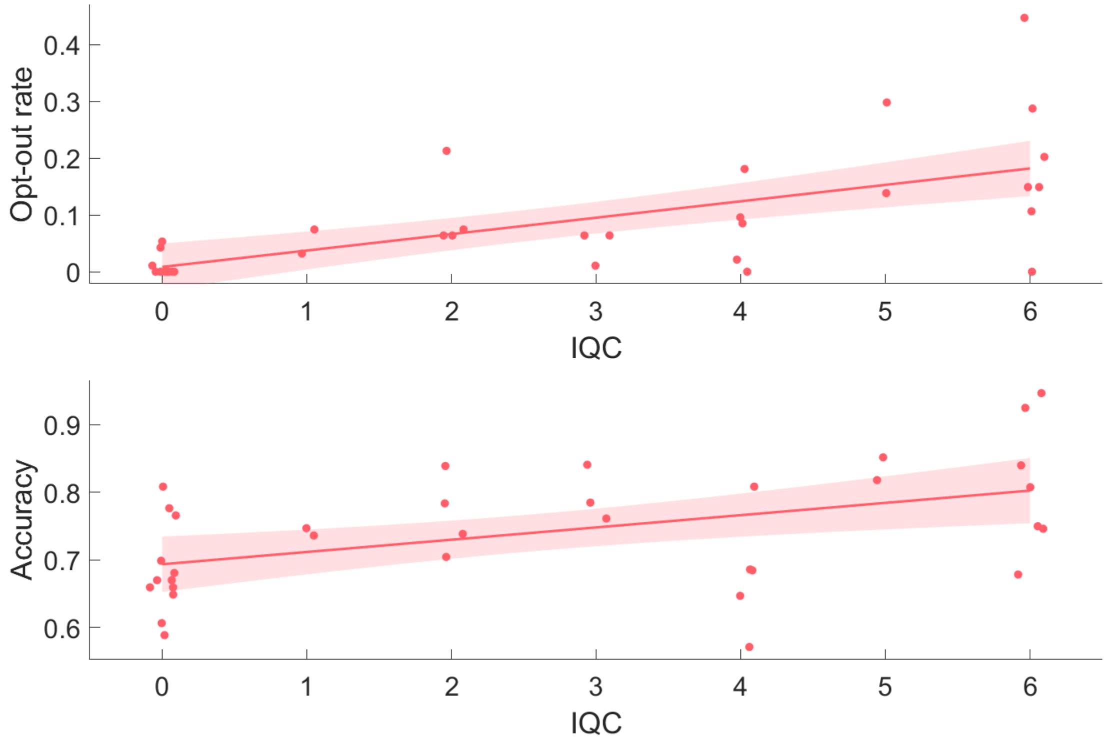


Figure 4.2: Opt-out rate and Accuracy on the 94 general knowledge questions in Experiment 1 as a function of the number of unanswered impossible questions (Impossible Question Criterion: IQC). Each point depicts a single participant with some random horizontal displacements for visual clarity. The line is the linear regression with the shaded region corresponding to the 95% confidence interval.

average accuracy of 95.1%, with only nine participants answering any incorrectly. We use these questions as an alternative method to filter, excluding those nine participants who answered any of the 6 easiest questions incorrectly. We assess crowd accuracy only on the 88 questions not used to make the filter to test the utility of these questions on identifying accurate respondents. The resulting 26 person crowd significantly outperforms the unfiltered crowd in terms of Proportion Correct (75.0 vs 71.7%, BF = 379.6), but not Proportion Better (60.6%, BF = 1.1).

While the crowd composed of participants with perfect accuracy on the easiest questions outperforms the unfiltered crowd, the impossible questions are a more restrictive filter. We compare the crowd composed of participants with perfect accuracy on the easiest questions to the impossible-question filtered crowd. We find that the impossible questions are a better filter for improving crowd performance measured by Proportion Correct (73.3% vs 79.7%, BF = 103.6) and Proportion Better (23.9%, BF = 34,933).

The 6 easiest questions and the 6 impossible questions are not unusual in their capacity to select for expertise and thereby improve crowd quality. Indeed, most combinations of 6 questions, when used as a filter, improve the quality of of the crowd. We sampled 10,000 random combinations of 6 general knowledge questions and used them to filter out participants as above. We created crowds composed only of those participants who answered all 6 of the randomly selected questions correctly. These crowds were evaluated using Proportion Correct on the 88 questions not used as a filter. The distribution of performance for these 10,000 crowds can be thought of as a null distribution that allow us to determine the extent to which a selection based on impossible questions in particular leads to improved crowd performance (see Figure 4.3)[2]. In terms of average Proportion Correct, the crowd made by filtering with Impossible Questions performs in the top 32.6% of crowds while the crowd formed with easy questions performs in the bottom 25.0%.

Table 4.2: Crowd performance in Experiment 1 depending on the Impossible Question Criterion (IQC) used to create crowds. Proportion Better and all Bayes Factors (BF) are computed in comparison to the crowd with all participants (denoted with an IQC of 0).

IQC	N	Proportion Correct (%)	BF	Proportion Better (%)	BF
0	35	73.2			
2	21	76.05	84.3	64.89	8.5
4	14	75.46	0.9	60.64	1.1
6	7	80.70	5177.2	70.21	318.2

4.3 Experiment 2

Experiment 1 demonstrated the potential of individuals' ability to identify and skip impossible questions as an indicator of metacognitive ability and as a method to compose more accurate

[2] It may be surprising to see that the impossible-question filtered crowd has variable performance. This occurs because the crowd composed in this fashion is evaluated on the same set of questions as the crowd filtered on a random set of 6 questions. This leads to 88 questions not used to filter either crowd, which vary randomly as the 6 questions used as a filter vary randomly.

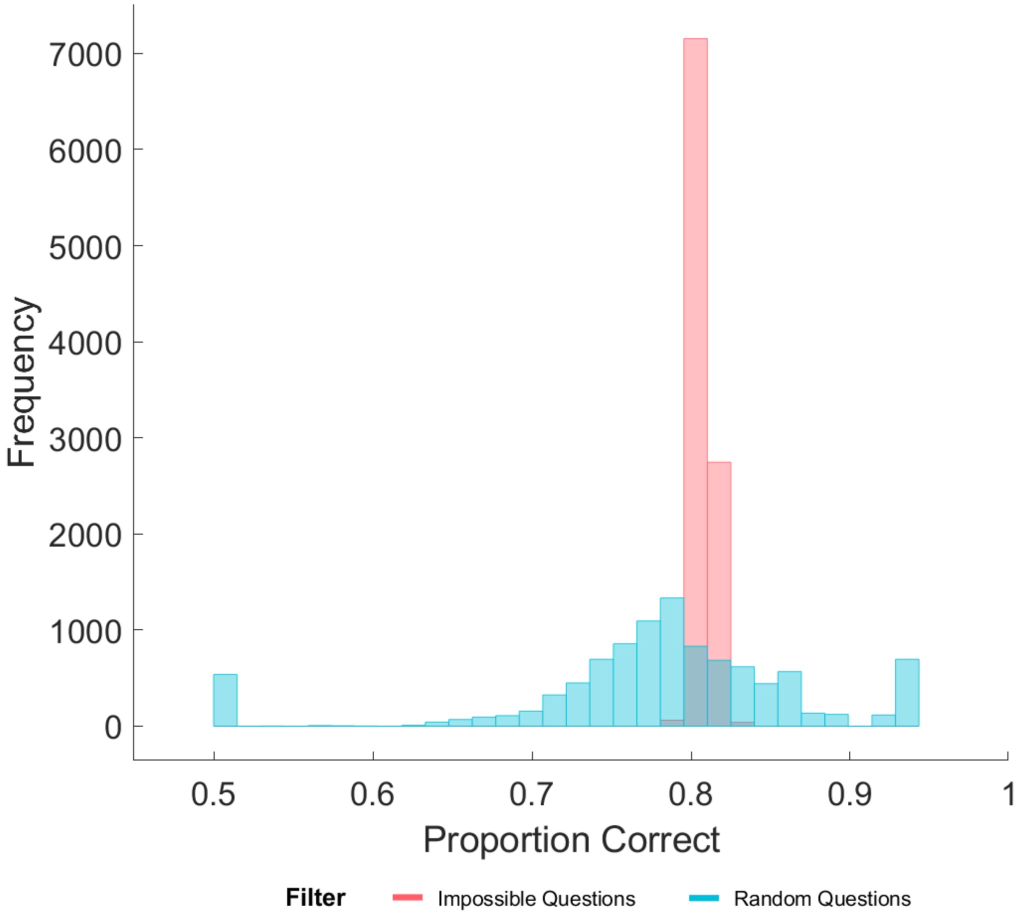


Figure 4.3: Proportion Correct for crowds composed of participants who correctly withheld answers from all 6 impossible-questions (red) or correctly answered 6 randomly selected questions (blue).

crowds. The reported effects are dependent on the responses of a small group of individuals with high metacognitive ability and so may be sensitive to the idiosyncratic responses of small groups. Experiment 2 addresses concerns of reliability by fully replicating the experimental design from Experiment 1 to examine whether crowds filtered with impossible questions *consistently* outperform those created using other filters.

4.3.1 Method

The methods for Experiment 2 were nearly identical to those of Experiment 1. The only difference in design was the number of participants recruited, with 32 participants recruited in Experiment 2 compared to the 35 recruited in Experiment 1.

4.3.2 Results

Data Overview

Participant answers are shown in Figure 4.4. Participants had an overall accuracy of 72.4% on the 94 general knowledge questions and opted in to those questions 95.6% of the time. Average confidence was higher when participants opted in than when they opted out (81.1% vs 54.6%, t = 12.1, BF $>$ 100,000).

Impossible Questions

Participants opted out of the 6 impossible questions much more frequently than they opted out of the general knowledge questions (4.4% vs 25.5%, BF $>$ 100,000 via Bayesian Contingency Table). The rate at which individuals opted out of impossible questions was related to a host of other behaviors. The participants who opted out of more impossible questions were also more accurate overall (r = 0.58, BF = 53.2) and more likely to opt out of general knowledge questions (r = 0.66, BF $>$ 100,000). Figure 4.5 shows these relationships.

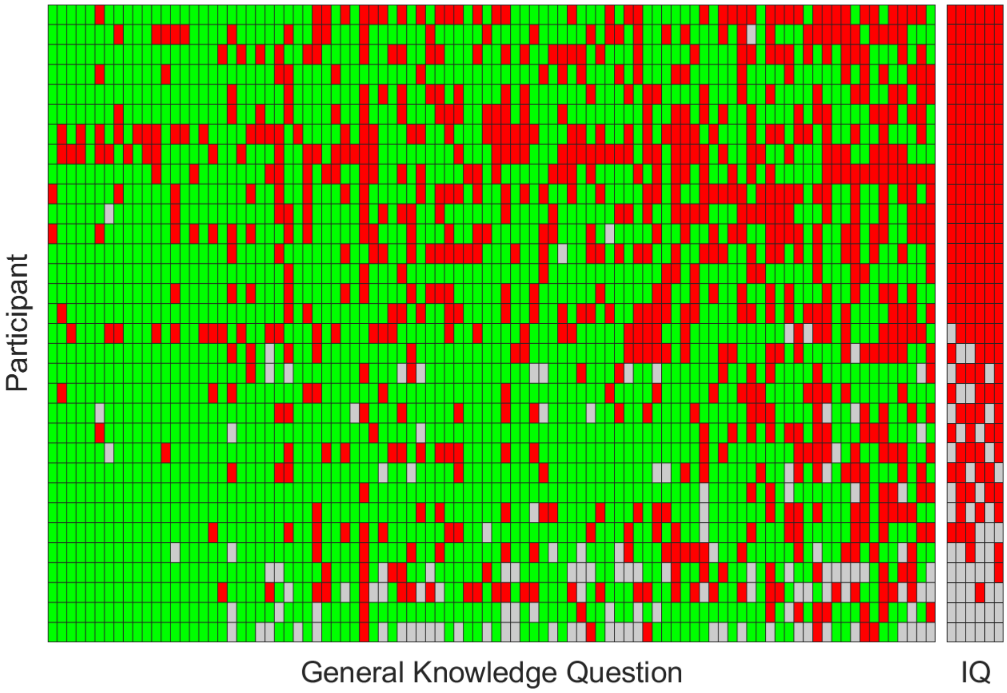


Figure 4.4: Participant responses to each question in Experiment 2. Green and red colors indicate a correct and incorrect response respectively. Grey colors indicate that the participant opted-out. Questions are sorted from left to right by increasing question difficulty as established by a pilot study. Impossible questions are labeled "IQ". Participants are sorted by the number of impossible questions answered.

Filters for improving crowds

We replicate the analyses from Experiment 1 for filtering and evaluating crowds. We filter out all crowd members who answer any of 6 questions incorrectly (impossible questions, the easiest questions, or randomly selected questions) and then assess the resulting crowd's performance. Using impossible questions as a filter in this way improves crowd performance relative to an unfiltered crowd (see Table 4.3).

In Experiment 2, participants had an average accuracy of 91.6% on the 6 easiest questions and 11 participants answered any of them incorrectly. We use these questions as a filter, excluding those 11 participants. In order to assess the ability of these questions to filter for good respondents, we assess crowd accuracy only on those questions not used to make the filter. The resulting 21 person

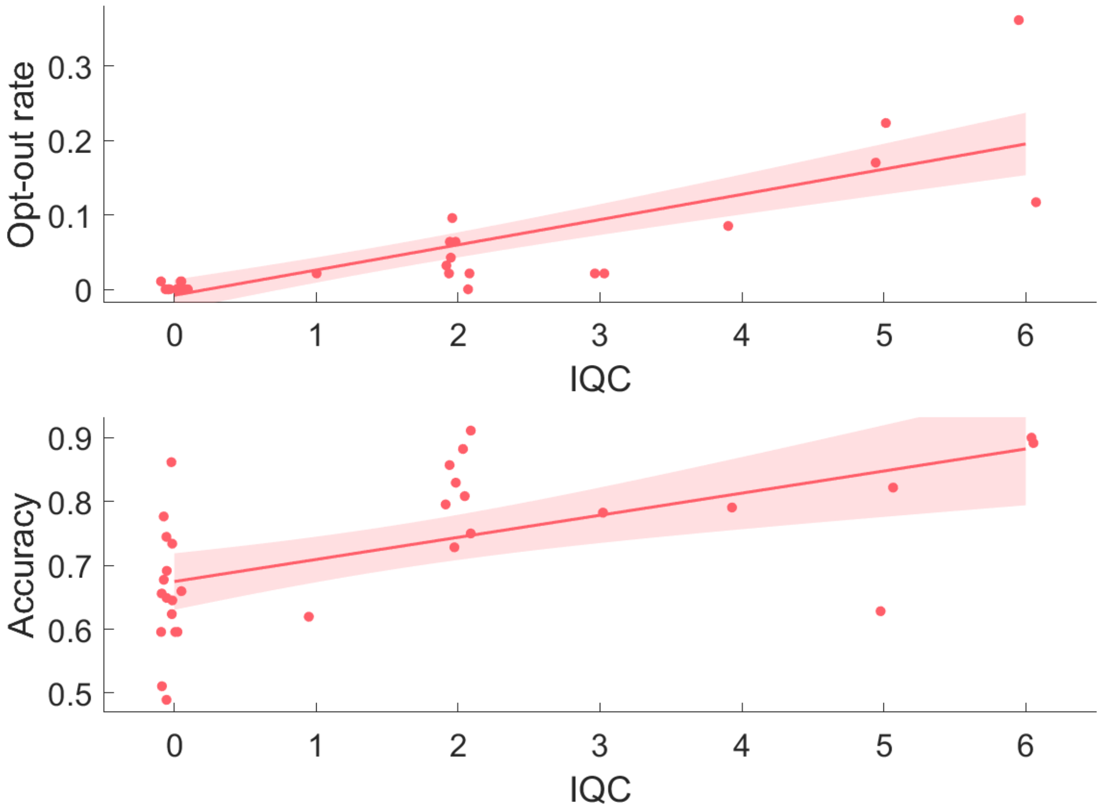


Figure 4.5: Opt-out rate and Accuracy on the 94 general knowledge questions in Experiment 2 as a function of the number of unanswered impossible questions (IQC). Each point depicts a single participant with some random horizontal displacements for visual clarity. The line is the linear regression with the shaded region corresponding to the 95% confidence interval.

crowd significantly outperforms the unfiltered crowd on the other 88 general knowledge questions in terms of Proportion Correct (73.5 vs 70.7%, BF = 268) and Proportion Better (73.9%, BF = 3,997).

As in Experiment 1, we sampled 10,000 combinations of 6 randomly selected general knowledge questions and used them to create crowds composed only of those participants who answered all of them correctly. Using this set of 10,000 crowds, we estimate a null distribution for each of the crowd performance metrics. These null distributions allow us to determine the extent to which impossible questions and very easy questions are effective filters for improving crowd performance (see Figure 4.6). The crowd made by filtering based on impossible questions performs in the top 2.1% of crowds while the crowd made using easy questions performs in the top 69.0%.

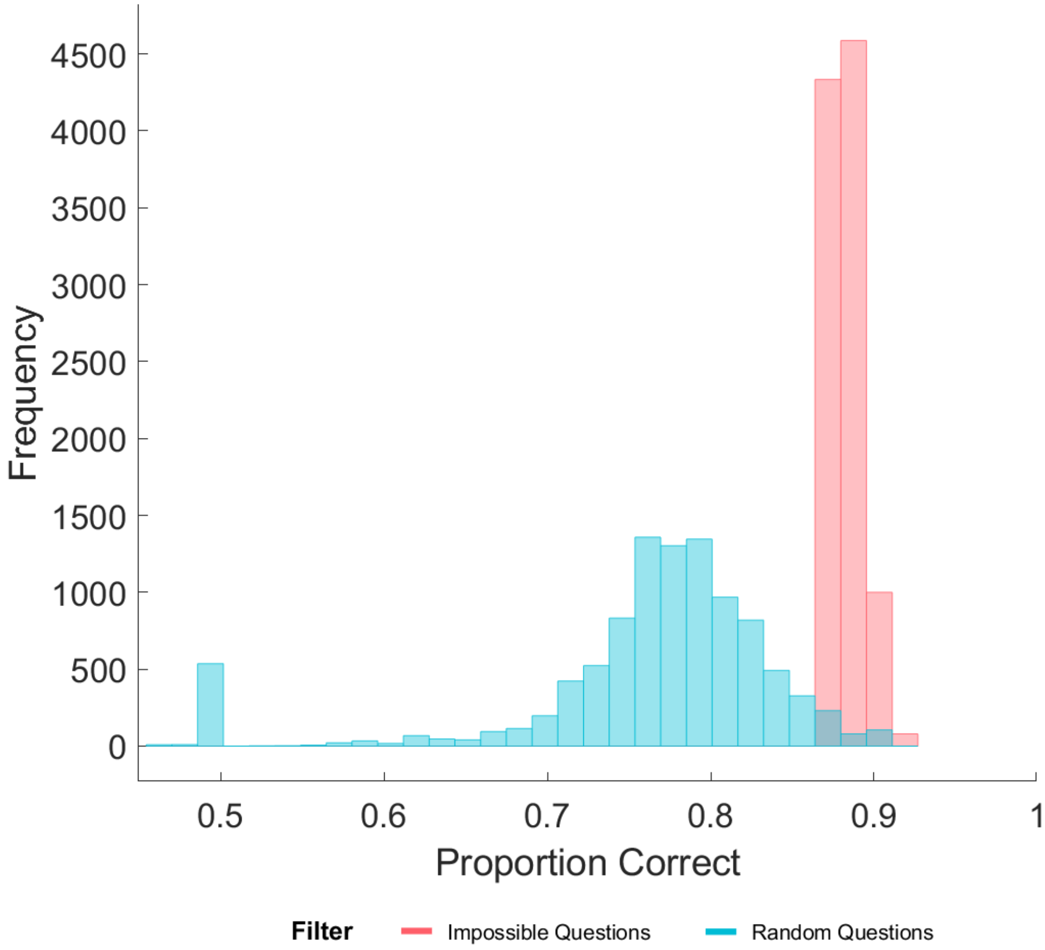


Figure 4.6: Proportion Correct for crowds composed of participants who correctly withheld answers from all 6 impossible-questions (red) or correctly answered 6 randomly selected questions (blue).

Table 4.3: Crowd performance in Experiment 2 depending on the Impossible Question Criterion (IQC) used to remove participants from the crowd. All analyses are compared to the unfiltered crowd (an IQC of 0), which includes all participants.

IQC	N	Proportion Correct (%)	BF	Proportion Better (%)	BF
0	35	73.2			
2	15	80.1	$> 100,000$	89.4	$> 100,000$
4	5	77.8	3.4	66.0	15.9
6	2	88.5	$> 100,000$	86.2	$> 100,000$

4.3.3 Discussion

We replicate our finding that impossible questions provide a measure of metacognitive ability which correlates with overall performance. Therefore, these questions can be leveraged to create high quality crowds. The performance of IQC as an indicator of the most useful crowd members persists despite the fact that very few participants correctly identified their lack of knowledge for all impossible questions. Traditionally, crowd wisdom is thought to result from large numbers of judgments averaging to the truth. Here we demonstrate that a better accurate method of achieving systematically high accuracy relies on the select few individuals that exhibit high metacognitive ability.

4.4 Reanalysis of Bereby-Meyer et al., 2003

[12] evaluated how scoring rules framed in terms of gains or losses impact test-taking strategies. The test allowed participants to omit (i.e. opt-out of) any number of multiple choice questions on a test that included 34 “solvable” items and 6 “unsolvable” items with no correct answers (i.e. impossible questions). Consistent with Prospect Theory, they found that framing the scoring rule in terms of gains instead of losses caused participants to be more cautious and answer fewer of both the solvable and unsolvable items.

This data provides an additional opportunity to examine how well metacognitive ability, as measured by impossible questions, predicts an individual’s performance on the task and their contribution to the crowd. Does the crowd composed of individuals who omitted the unsolvable items outperform crowds that include all participants?

In this experiment, they recruited 92 participants from Ben-Gurion University in Israel. Each participant answered 34 general knowledge questions covering topics such as geography, history, and art. Each question had four possible answers, only one of which was correct. Each participant was also asked 6 impossible questions with no correct answer. Accuracy was incentivized by providing extra course credit for the participants who scored in the top 50%. To our knowledge, there was no overlap in the questions asked in this study and the questions asked in Experiments 1 and 2.

4.4.1 Results

Average accuracy in their experiment was 58.2%. As in our experiment, participants opted-out of the general knowledge questions less frequently than the impossible questions (13.6% vs 29.4%, BF > 100,000). Only 40 of the 92 participants opted out of any impossible question and 8 of them opted out of all 6. IQC was positively correlated with accuracy (r = 0.33, BF = 25.1) and the rate at which participants opted-out of the general knowledge questions (r = 0.86, BF > 100,000).

As before, we examine the performance of the crowd composed of participants who opted in to different numbers of impossible questions and compare it to the performance of a control crowd which uses all participants (see Table 4.4). Filtering out respondents that exhibit low metacognitive ability by enforcing a minimum IQC generally results in better crowd performance in terms of both Proportion Correct and Proportion Better.

Table 4.4: Crowd performance for [12] depending on the Impossible Question Criterion (IQC) used to remove participants from the crowd. All analyses are compared to the unfiltered crowd (an IQC of 0), which includes all participants.

IQC	N	Proportion Correct (%)	BF	Proportion Better (%)	BF
0	92	57.2			
2	35	61.9	282.7	73.5	9.4
4	28	62.5	50.5	76.5	27.0
6	8	65.8	0.7	62.5	0.5

4.4.2 Discussion

We replicate our core findings involving a different set of impossible questions: they identify high performers and as a result are useful for forming accurate crowds. These results indicate that impossible questions are a reliable indicator of expertise across experimental contexts.

4.5 General Discussion and Conclusion

An individual's propensity to skip impossible questions is easy to assess and readily identifies high performers. We demonstrated that impossible questions can be used to identify experts and form highly accurate crowds in a novel experiment, its replication, and in a reanalysis of data from [12]. Moreover, filters based on impossible questions outperformed most other sets of possible control questions in identifying experts, demonstrating that metacognitive ability can be a better predictor of expertise than historical accuracy.

As a measure of expertise, impossible questions can be especially useful in contexts where expertise is difficult or costly to evaluate. In forecasting contexts, direct measures of participant ability such as Contribution Weighted Scoring [18] can only be used after forecasters have an established history and the true outcomes to several questions are known. Because the "truth" of impossible questions is known ahead of time, impossible questions can provide a measure of expertise as soon as a participant joins the platform. In recommender systems, the difficulty of getting early measures of item difficulty and member expertise is called the *cold-start problem*. Impossible questions may allow for an early estimate of individual ability, adding to existing techniques that address the cold-start problem for item difficulty [102].

Much of the existing work relying on the efficacy of crowds grants the experimenter complete control over question selection. Our setup is fundamentally different in that participants choose which questions to answer for themselves. As a result, each observed response in the crowd has passed through a metacognitive filter, and so the quality of that filter is of interest. Many real-world crowd-sourcing platforms share this feature of participant choice (e.g. Predictit, The Good Judgment Project, Wikipedia). Metacognition is a valuable but under-explored area of research because it relates to crowd wisdom in real-world applications.

While impossible questions may be a useful measure of ability in situations that rely on participant choice, they are not the only measure of metacognition. In contexts more suitable to their measurement, meta-d' or other metacognitive measures may be useful indicators of expertise in contexts that allow participants to opt-in (for a comparison of Impossible Questions to confidence, see the Appendix). Indeed, many existing methods that improve the accuracy of crowds implicitly rely on metacognition. Confidence-weighted pooling exploits the relationship between the metacognitive judgment of confidence and item-level accuracy [79], Bayesian Truth Serum exploits the relationship between meta-knowledge of others' beliefs and accuracy [84, 85], and the benefit of opting-in is due to participants' ability to recognize their own knowledge [10]. Making this connection explicit highlights the need for further research on role of metacognition in crowd contexts.

Chapter 5

SAGE: A Geopolitical Forecasting Case Study

In 2017, the organization Intelligence Advanced Research Projects Activity (IARPA) hosted a Hybrid Forecasting Competition that lasted through the beginning of 2020. The goal of the project was to develop strategies to leverage a hybrid of human forecasts and algorithmic forecasts in order to predict the outcome of geopolitical events. This chapter is a case study detailing my contributions to the Synergistic Anticipation of Geopolitical Events team (SAGE) [75] that touch on the research topics contained in this dissertation. The two questions at the heart of this case study are: how should human and algorithmic forecasts be combined, and to what extent should humans self-direct in a forecasting context?

The forecasting competition provided interesting challenges that differentiate it from a traditional laboratory study. Three teams composed of behavioral scientists and machine learning researchers competed to accurately forecast the outcomes of future geopolitical events. Performance was evaluated relative to a Control team run by IARPA that used state of the art methods similar to those of the Good Judgment Project to collect and aggregate forecasts from non-expert forecasters with opinion pools [104, 108]. Competing teams were tasked with improving upon this benchmark by developing their own platform for eliciting forecasts, aggregation strategies for combining these forecasts, and machine learning algorithms to automatically forecast on questions. These machine learning algorithms were the primary advantage of the competing teams and the goal of the competition was to evaluate promising methods of creating human-algorithm hybrids that outperform the human-only Control. Questions were selected by the MITRE corporation from a variety of geopolitical topics such as international security and global health. Since questions were not generated by the competing teams, methods developed for the competition reflect generalizable strategies for forecasting. Methods could not be tailored to the specific questions asked in the competition since they had to be specified prior to the release of any questions. Our team's performance on these questions was evaluated starting on the day the questions were released and so successful methods of generating forecasts needed to arrive at accurate forecasts quickly even when information was sparse. The interdisciplinary nature of the project allowed machine learning researchers and behavioral scientists to collaboratively design methodologies for addressing these challenges.

One benefit of forecasting competitions is that they fulfill a role similar to that of preregistration. A typical pipeline for psychological research involves the collection of data, evaluation of various models on that data, and then publication of the best performing model. The published model might be a simple account of the data, such as a non-zero correlation between two variables via a t-test; or otherwise might provide a more involved account of the data via e.g. a Bayesian cognitive model. In either case, the researcher will frequently have considered many models in the course of data analysis. Indeed, it is a recommended practice to quickly iterate through many possible models and while there are strategies to avoid overfitting from post-selection inference [34], usually only the best performing model is recorded in publication. As a result there exists a survivorship bias; models will typically only be observed in publication when they perform well on the the statistical measure used in model evaluation. In the case of performing t-tests to evaluate many possible independent variables that have a non-zero correlation with an outcome of interest, this is sometimes called multiple hypothesis testing or p-hacking, and results in spurious findings.

Sophisticated models have even more degrees of freedom. While cross-validation can mitigate the degree to which these more sophisticated models overfit [14], it cannot prevent it altogether due to the iterative nature of data analysis wherein candidate models are repeatedly evaluated on the test data [87, 25]. This competition, in contrast, required that models be specified prior to data collection and so each model's performance was a genuine reflection of its ability to predict that data.

5.1 Project Overview

Competing teams were tasked with forecasting the outcomes of hundreds of geopolitical questions over the course of the competition. Questions covered a variety of topics including International Relations, Global Health, Economics, and Technology. Questions would be released by MITRE throughout the competition and available for forecasting for between 5 and 200 days until the outcome became known and performance was evaluated. Example questions are shown in table 5.1. This analysis will focus on the second phase of the tournament that spanned the last 9 months of the competition and included a total of 399 questions. Each competitor was allowed to specify 100 aggregation strategies prior to the start of the competition and each competitor's performance was based upon their best performing aggregate.

5.1.1 Participants

We were provided with 600 workers from Amazon Mechanical Turk's "Turk Prime" each week. Most workers chose to return each week, with those who chose not to return being replaced with new workers from Turk Prime. These workers logged onto our platform and provided 5 forecasts on various geopolitical questions. Forecasts took the form of opinion polls; probability estimates for each of the possible outcomes that summed to 100%. Of the 5 forecasts they were required

Table 5.1: Example questions from the Hybrid Forecasting Competition, correct answers are shown in bold.

Question	Response Options
Before 30 November 2019, will Libya set a date or dates for parliamentary and/or presidential elections?	(A) Yes, both parliamentary and presidential (B) Yes, only parliamentary (C) Yes, only presidential **(D) No date for either presidential or parliamentary elections will be set before 30 November 2019**
Will the Government of Canada issue a travel advisory of "Exercise a high degree of caution," "Avoid non-essential travel," or "Avoid all travel" for New Zealand between 4 April 2019 and 12 June 2019?	(A) Yes **(B) No**
What will be the Korean Central News Agency (KCNA) Threat Index on 10 June 2019?	(A) Less than 0.2 **(B) Between 0.2 and 0.4, inclusive** (C) More than 0.4 but less than 0.6 (D) Between 0.6 and 0.8, inclusive (E) More than 0.8

to provide, 2 were on questions they had not yet answered. The other 3 forecasts were updates to previous questions, either revising or affirming their forecast. Forecasters were encouraged to research the subject and provide a rationale for their prediction.

When viewing questions, forecasters were able to see the rationales and forecasts of other forecasters. When available, forecasters were also shown historical data for the question at hand. For instance, the question "What will be the Korean Central News Agency (KCNA) Threat Index on 10 June 2019?" was accompanied by a chart depicting the Threat Index reported by the KCNA for the past 6 months. In addition, a subset of the forecasters were shown automatically generated algorithmic forecasts based on this time series. Forecasters were provided training materials explaining how to conduct research, generate probabilistic forecasts, and interpret algorithmic forecasts.

5.1.2 Scoring

Competitors' accuracy was evaluated using the Brier score and Cohen's d metrics. These scoring rules were determined by IARPA prior to the start of the competition. The Brier score is usually equivalent to the squared error of the forecast[1]. Competitors' aggregate forecasts were scored for each question on each day, and the average performance of questions was termed the Mean of the Mean Daily Brier score (MMDB). Equation 5.1 details this computation based upon the Brier score (B) for each forecast (f) relative to the outcome (o) across all questions (N) and all days (T). Lower Brier scores correspond to better performance. This metric is a *proper scoring rule* and so forecasters can minimize their expected Brier score by reporting their true beliefs for each question on each day. It is important to note that because the score on a question was equally weighed across all days it was available, quickly arriving at accurate forecasts conferred a significant advantage.

$$MMDB = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{j=1}^{T}B(f_{ij}, o_i) \tag{5.1}$$

Cohen's d measured the extent to which a competitor systematically outperformed the Control condition run by IARPA. Cohen's d was computed according to equation 5.2, where $MMDB_{condition}$ corresponds to the Mean of the Mean Daily Brier score of the referenced condition and *SD* corresponds to the Standard Deviation. Higher values of Cohen's d correspond to better performance.

$$d = \frac{MMDB_{competitor} - MMDB_{control}}{SD(MDB_{competitor} - MDB_{control})} \tag{5.2}$$

Strategies that produce optimal forecasts under the Brier score will not necessarily maximize Co-

[1] The Brier score was modified to accommodate ordinal questions so that forecasts with probability mass close to the true outcome resulted in a lower Brier score than a forecast with probability mass distant from the true outcome. The KCNA threat index question is an ordinal question with the second response option being correct, so a forecast of [0.5,**0.5**,0.0,0.0] would have a lower Brier score than a forecast of [0.0,**0.5**,0.0,0.5]. This is because in the former forecast the probability mass that did not fall into the correct response option is closer to the true outcome.

hen's d. For example, consider a forecaster who generates well-calibrated forecasts on binary choice questions with confidence varying uniformly from 50-100%. This forecaster can minimize their expected Brier score by reporting their true beliefs because the Brier score is a proper scoring rule, which would result in an expected Brier score of 0.33. If the control forecaster used in calculating Cohen's d always reported complete uncertainty by providing 50% confidence in each response option, then the well-calibrated forecaster would have an expected Cohen's d of 0.42. However, this well-calibrated forecaster could improve their expected Cohen's d by reporting a forecast that is the mean of their genuine beliefs and the uniform forecast reported by the control forecaster. In expectation, this strategy would result in less accurate forecasts as measured by Brier score (0.38) but an improvement in Cohen's d (0.58). In this sense, Cohen's d is not a proper scoring rule since a forecaster can improve their performance by reporting something other than their true beliefs. These scoring rules were chosen by IARPA prior to the start of the competition to encourage accurate forecasts that improve upon the baseline.

5.2 Aggregation

Aggregation is the task of using statistical techniques to combine a collection of forecasts with the goal of arriving at the most accurate forecast possible. *Hybridized* aggregation incorporates judgments from both humans and machine learning algorithms. Both humans and algorithms have comparative advantages when forecasting that can be leveraged in aggregation to generate forecasts that outperform either component individually.

Human forecasters are capable of independent research and so analyze both quantitative and qualitative data to reason about likely future outcomes. While individual human forecasters may exhibit biased forecasts, large and diverse crowds of forecasters can be combined with opinion pools or prediction markets to form highly accurate predictions of the future, termed the Wisdom of Crowds effect [1, 76, 104].

Algorithms are capable of automatically forecasting on large numbers of questions and updating to reflect incoming data. This makes algorithms appealing when scaling to large numbers of questions. Further, algorithms present the opportunity for a forecasting source that interprets available data in a different manner than human forecasters. To the extent that algorithms represent a signal of the ground truth that has differs from human forecasters, the algorithms can be used to improve aggregate accuracy relative to a human-only baseline.

Since each team was permitted 100 aggregation strategies, our team developed and deployed strategies that varied the method of aggregating human judgments and the method of creating human-algorithm hybrids. We were permitted to use the Control forecasters in aggregation and so we also varied the source of human forecasts, including strategies that use SAGE forecasters only, Control forecasters only, or a combination of SAGE and Control forecasters. Each of these aggregation strategies had to be fully specified prior to the start of the competition and could not be altered once the competition began. We were permitted to use adaptive strategies that updated parameter values so long as the algorithm to update parameter values was fully described before the competition began.

This section will focus on the hybridized aggregation strategies our team employed in the competition. The effects of hybridization vary slightly depending on the human forecasts used and the method of aggregating human forecasts. While the different aggregates of human forecasts resulted in similar forecasts to one another, only the best performing human aggregation strategy will be reported in order to isolate the effects of hybridization in the ideal case.

5.2.1 Human Aggregation

Human forecasts were aggregated with a weighted opinion pool that used statistical corrections for sources of human miscalibration. These statistical techniques were developed by other members of the SAGE team and built upon known methods [1]. Forecasts were weighed based on their historical accuracy according to their contribution weighted score [21] so that forecasters who have historically improved aggregate accuracy were given higher weight. Forecasters were also given higher weight for giving rationales that included numbers and for a tendency to provide small regular updates in light of recent news (as opposed to large updates or no changes whatsoever). Since geopolitical forecasting involves incorporating time-sensitive news and data, only the most recent 40% of forecasts on a question were included in the aggregate. Further, exponential decay was applied to downweight older forecasts. Human forecasts tend to be overly extreme (closer to 0% or 100%) and yet the average of these forecasts tends to be underextreme. Both of these biases were corrected for and collectively resulted in aggregate forecasts that were slightly more extreme than they would be without any correction.

5.2.2 Algorithmic Forecasts

Other members of the SAGE team created algorithms to automatically generate forecasts based on either the historical data or the text of the question. For questions with quantitative historical data, forecasts were generated using a variety of time series models, including univariate and multivariate auto-regressive moving average models (ARIMA and ARIMAX). Specialized models were also included to forecast on geopolitical events that could be coded using the CAMEO framework [35]. Questions with no historical data were forecasted with a model that assumes that questions have a similar rate of status-quo outcomes as its k-nearest neighbors based on a language model of the question text. In all cases where multiple algorithmic forecasts were available, an ensemble with feature selection based on Discrete Cosine Transform was employed to combine the algorithmic forecasts. In total, algorithmic forecasts were available for 299 of the 399 questions in the tournament. For a comparison of these algorithms with the human forecasters in a previous phase of the competition, see [4].

5.2.3 Hybridized Aggregation Methods

Two general approaches were used to combine human and algorithmic forecasts into hybrid aggregates. In the first (M1), the hybrid aggregate was a linear average of the the human aggregate and the algorithmic forecast. In the second (M2), the algorithm was treated as equivalent to approximately six human forecasters, so as more humans answered a question the aggregate became increasingly comprised of those humans' forecasts.

M1 aggregate forecasts (F_{M1}) were generated according to equation 5.3 by computing a linear average of the aggregate human forecast (F_{human}) and algorithmic forecast ($F_{algorithm}$) based upon algorithmic weight α,

$$F_{M1} = \alpha F_{algorithm} + (1-\alpha) F_{human} \tag{5.3}$$

M2 aggregate forecasts (F_{M2}) were equal to an average of all human forecasts and the algorithmic forecast, each termed F_i, and weighed according to learned weights α_i,

$$F_{M2} = \frac{1}{\sum_{i=1}^{n} \alpha_i} \sum_{i=1}^{n} \alpha_i F_i \tag{5.4}$$

Weights for human forecasters were learned according to the human aggregation strategy and were normalized to have a mean of 1. As a result, the M2 algorithmic weight, α, was the average number of human forecasters that would be required to create an aggregate that is equal parts human and algorithm. It can be thought of as the number of human forecasts the algorithm was "worth" in aggregation.

Both M1 and M2 were adaptive algorithms that learned algorithmic weights as question outcomes became known. Algorithmic weights were chosen to minimize the Brier score of the questions with known outcomes with L2 regularization used to prevent overfitting. The details of this calculation are shown in equation 5.5 where α is the learned weight, *MMDB* is the Mean of the Mean Daily Brier score, $F_M(\alpha_*)$ are the forecasts generated by model M1 or M2 with $\alpha = \alpha_*$, and α_0 is the prior weight. Prior weights were specified by selecting the α values that would have minimized the Brier score in a previous phase of the tournament, with $\alpha_0 = 0.1$ for M1 and $\alpha_0 = 6$ for M2. λ governed the rate at which the algorithm weights were updated where $\lambda = 2$ for M1 and $\lambda = 1$ for M2.

$$\alpha = \arg\min_{\alpha_*} MMDB(F_M(\alpha_*)) + \lambda(\alpha_0 - \alpha_*)^2 \tag{5.5}$$

Relative to M1, M2 gave higher weight to the algorithm when the total human weight was low. This would occur when there were few human forecasters, human forecasters were of low average expertise, or the human forecasts were outdated. When there was a large and highly expert crowd of human forecasters, M2 aggregate forecasts were more similar to the human-only aggregate than M1. Interpreting M2 in a Bayesian light, the algorithmic forecast can be thought of as an informative prior from which each human forecast is an update with strength proportional to the forecaster's weight.

5.2.4 Hybrid Aggregate Performance

Aggregate model performance was evaluated using the Brier score and the Cohen's d metrics on the 299 questions with algorithmic forecasts. Results for all combinations of hybridized models and human inputs are reported in table 5.2. For all sets of human inputs (Control, SAGE, or Control & Sage), both hybridization models (M1 and M2) improve aggregate performance, with M2 conferring the largest benefits to forecasting accuracy in terms of observed Brier score. Indeed, M2 combined with the best performing human aggregation algorithm and applied to both SAGE and

Control forecasters resulted in the most accurate forecasts of all aggregation strategies deployed by any team in the competition. However, M1 applied to only the Control condition resulted in the hybrid aggregate forecasts with the strongest evidence of some benefit over the Control, measured by both the Bayes factor and Cohen's d. Therefore, the gains from M1 may be more reliable than the gains from M2. These improvements in aggregate accuracy by incorporating algorithmic forecasts with M1 and M2 occur despite the algorithmic forecasts performing significantly worse than the control aggregate on average.

Table 5.2: Aggregate model performance depending on both the hybrid model and source of human forecasts (human inputs). M1 takes a linear average of the human aggregate and the algorithmic forecast. M2 treats the algorithm as a human forecaster with weight equal to approximately 6 humans. Model performance was evaluated only for the questions with algorithmic forecasts available (N=299). MMDB is the Mean of the Mean Daily Brier score and Cohen's d is the effect size of the mean daily Brier score relative to the human-only control aggregate. Bayes factors (BF_{10}) were computed via a two-tailed paired samples t-test relative to the human-only Control using a Cauchy prior with scale equal to 0.707.

Hybrid Model	Human Inputs	MMDB	BF_{10}	Cohen's d
None	Control	0.327		N/A
M1	Control	0.315	2,364	0.271
M2	Control	0.309	3.5	0.165
None	SAGE	0.308	0.1	0.075
M1	SAGE	0.305	0.2	0.088
M2	SAGE	0.302	0.2	0.085
None	Control & SAGE	0.307	0.7	0.127
M1	Control & SAGE	0.299	5.6	0.175
M2	Control & SAGE	0.296	6.6	0.178
Algorithm Only	None	0.416	>100,000	-0.177

The final weights learned by the model deviated slightly from the weights set a priori. M1 weights, expressed as a proportion of the total forecast that stems from the algorithm, were lower than would have been optimal and the final learned weights ranged from 0.16 to 0.20 compared to the prior of 0.1. M2 weights, expressed as the number of forecasters the algorithm was "worth" were slightly higher than optimal and final learned weights ranged from 4.3 to 5.8 compared to the prior of 6.

To compare algorithmic weights between M1 and M2, M2 weights were converted into the pro-

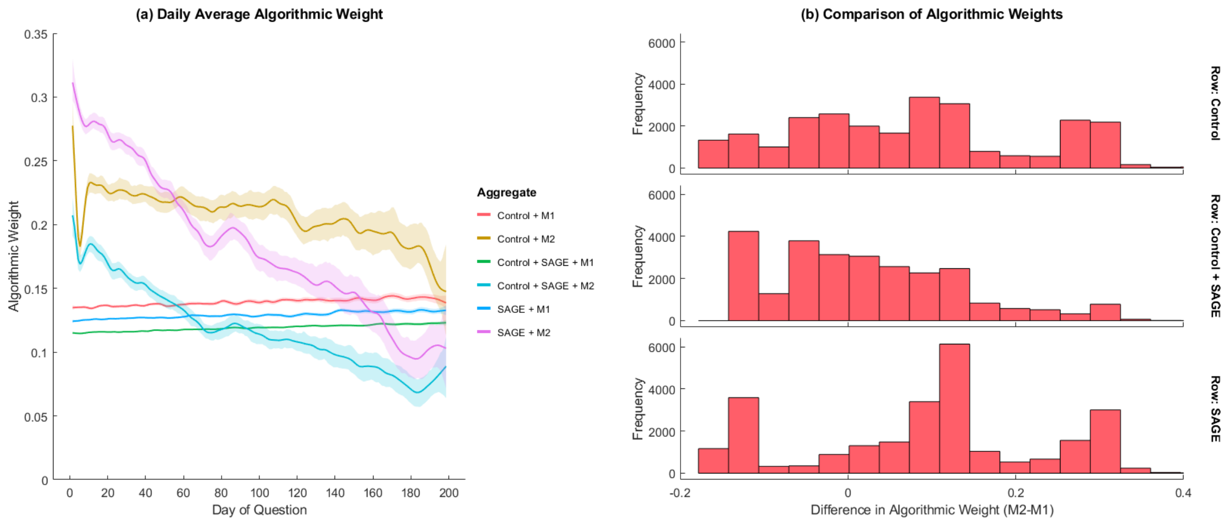


Figure 5.1: (a) Average algorithmic weight depending on how many days a question has been available for each aggregation strategy. The solid lines pass through the mean algorithmic weight on each day and the shaded regions represent the 95% confidence interval associated with that day. (b) The distribution of the difference in algorithmic weight between M1 and M2 for each set of human forecasters used in aggregation. Positive values correspond to higher weight assigned to the algorithm in M2 than M1. Algorithm weights tended to be higher under M2 than M1, although this difference in weights varied significantly.

portion of the aggregate forecast that stemmed from the algorithm. Recall that when using M1, algorithmic weight only changed as it was updated in light of past performance. In contrast, when expressing M2 algorithmic weight in these terms, the contribution of the algorithm varied depending upon the number of human forecasters, their estimated expertise, and the recency their forecasts. As a result, M2 algorithmic weights varied substantially both between questions and over time within questions. Algorithmic weights differed substantially between the two aggregates, although M2 weights were on average higher than M1 (see fig 5.1). Any differences between the forecasts generated by M1 and M2 were due entirely to these differences algorithmic weight.

5.2.5 Scalability

The gains of hybridized aggregation observed in this project were moderate. However, hybridized aggregates have the potential to scale to larger numbers of questions than human-only aggregates.

When the algorithm provides a reasonable baseline forecast, fewer human forecasters may be needed to arrive at an accurate aggregate than when humans are the only source of forecasting.

Scaling to larger numbers of questions would correspond to each question receiving fewer forecasts. A simulation was conducted to model the effect of scaling to larger numbers of questions by randomly censoring forecasts from each of the aggregation models. In the simulation, 10% to 90% of the human forecasts were censored at random N=20 times each and the performance of each aggregation model (human-only, M1, and M2) was recorded. Results from this simulation can be seen in figure 5.2 for both the Control and SAGE forecasters. To assess the degree to which incorporating algorithmic judgments conferred increasing improvements to Brier score as sparsity increased, a least-squares linear regression was used. As sparsity increased in the Control condition, the benefits of using hybridized aggregation strategies increased for both M1 (R^2 = 0.41, BF $>$ 100,000) and M2 (R^2 = 0.72, BF $>$ 100,000). This was also true when only considering forecasters from the SAGE condition for both M1 (R^2 = 0.60, BF $>$ 100,000) and M2 (R^2 = 0.57, BF $>$ 100,000). Sparse crowds received increasing benefits from hybridized aggregation.

Another way of interpreting the results from this simulation is to identify the number of forecasts that could be removed while maintaining the level of performance observed in the human-only aggregate. For instance, the human-only Control aggregate scored a MMDB of 0.328. By incorporating algorithmic judgments with M1, on average that can be achieved with only 30% of the forecasts (MMDB = 0.325) or with M2 and only 20% of the forecasts (MMDB = 0.319). When using SAGE forecasters, the human-only aggregate achieved a MMDB of 0.308. On average, this forecasting accuracy can be matched with 70% of the forecasts by using M1 (MMDB = 0.307) or 50% of the forecasts by using M2 (MMDB = 0.308). These results suggest that algorithmic forecasts can be used to maintain a desired level of accuracy even when far fewer forecasts are available and so reduce the required labor to accurately forecast the future.

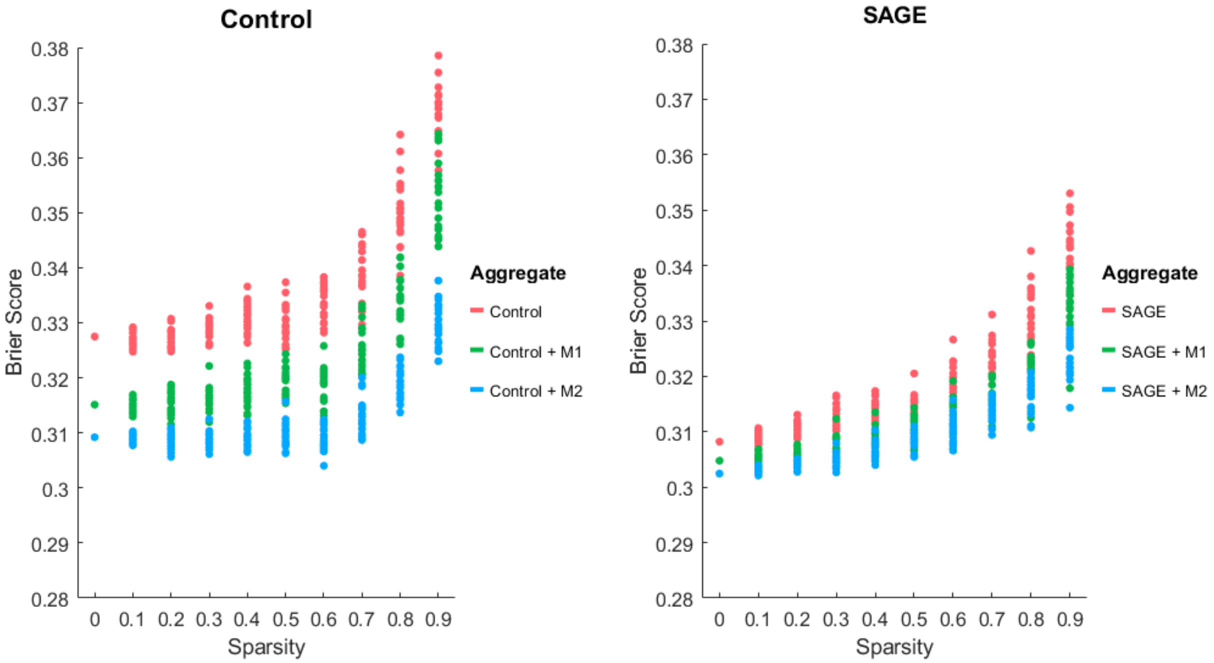


Figure 5.2: Aggregate model performance (Brier Score) depending on the proportion of human forecasts removed from the forecasting pool (Sparsity). For each level of sparsity from 10% to 90%, N=20 simulated sets of forecasts were censored from the aggregate.

5.3 Self-direction

When considering the issue of how much control forecasters should have over question selection, two effects suggest opposing courses of action. The first effect pressures the platform to allow forecaster choice because doing so allows forecasters to leverage their metacognitive abilities to choose questions for which they have relevant knowledge. Indeed, forecaster question selection is an important component of forecaster skill [70], with highly accurate forecasters selecting many questions across a wide range of topics. In the domain of general knowledge questions, permitting members of a crowd to opt-in improves that crowd's accuracy by allowing crowd members to choose questions for which they have relevant knowledge [10]. The second effect pressures the platform to restrict forecaster choice in order to satisfy the conditions under which aggregation improves the accuracy of crowds for all questions. These conditions are met when crowds are large and diverse [23], and so ensuring that forecasters disperse across a wide variety of questions should improve crowd accuracy. How should a platform manage these opposing forces and what degree of self-direction should forecasters be permitted?

In the case of general knowledge questions, the benefits of self-direction outweigh the costs in terms of crowd size and diversity [10]. However, there are some reasons to believe that those results may not generalize to the Hybrid Forecasting Competition. In the competition, forecasters were encouraged to research questions and so knowledge and meta-knowledge at the time of question selection may not have predicted their future performance.

Initially, our platform allowed forecasters to freely choose which questions to answer. Forecasters exhibited a preference for a relatively small subset of questions and seven weeks into the competition 54% of the forecasts concentrated in only 25% of the 111 questions available at that time. The control condition had a similar distribution of responses across questions, with 52% of the forecasts being concentrated in only 25% of questions. This density of forecasts and the limited labor available to us meant that many questions had "crowds" composed of as few as 5 forecasters even after being available on the platform for a full week.

To estimate the degree to which effort was being wasted on these popular questions, the effect of limiting each question to a fixed number of forecasts on aggregate performance was simulated. Various limits were tested wherein forecasts were randomly removed from each question that had received more forecasts than the limit until the limit was met. Aggregate performance was then evaluated on these simulated crowds of forecasters. The large number of forecasters on these very popular questions conferred minimal benefits to our overall Brier score on the few questions for which outcomes were known at the time (N=18, see table 5.3), and so the work associated with those forecasts would likely have been better spent on other less popular questions.

To make better use of the limited SAGE forecasters, our team introduced limited forced choice for question selection starting in week 8. Rather than selecting from the full list of all unresolved questions, forecasters selecting new questions were only presented the 10 questions with the fewest

Table 5.3: Aggregate performance (MMDB) as a function of a simulated limit on the number of forecasts allowed for each question. To simulate the limit, forecasts were randomly removed from each question until the limit was reached and aggregate MMDB was recorded. This analysis only includes the N = 18 questions resolved during the first seven weeks of the competition and was run 1,000 times to account for sampling error. MMDBs reported below correspond to the mean MMDB of the human-only aggregate applied to all simulated crowds.

Simulated Limit	Proportion Remaining	MMDB
75	49.2%	0.3806
150	87.6%	0.3796
∞	100%	0.3777

forecasters. The goal of this intervention was to permit a degree of choice that both captured any gains from forecasters' metacognitive judgments and and ensured that less popular questions would receive sufficient forecasts to benefit from aggregation.

After limiting forecaster choice in this way, forecasts were more evenly distributed across the available questions than they would have been without it (see fig 5.3). While forecasts were as unevenly distributed in the SAGE and control conditions before the intervention, by the end of the competition SAGE forecasts were more evenly distributed: the top 25% most popular questions received 40% of the SAGE forecasts but 47% of the Control forecasts.

In order to assess the impact of limiting forecaster choice on aggregate accuracy, the difference between SAGE and Control aggregate performance was observed both before and after the intervention (see table 5.4). Since some questions received forecasts both before and after the intervention, questions were binned into either "full" or "limited" choice depending on the number of days the question was available either before or after the intervention. Questions that were available for more days while forecasters had full control over question selection were considered to have "full" forecaster choice while those that did not were considered "limited". In either case, only the forecasts that occurred before or after the intervention were included in the "full" or "limited" aggregates respectively. To test whether or not limiting choice in this way represents a statistically reliable method of improving aggregate accuracy, a Bayesian t-test was computed to evaluate

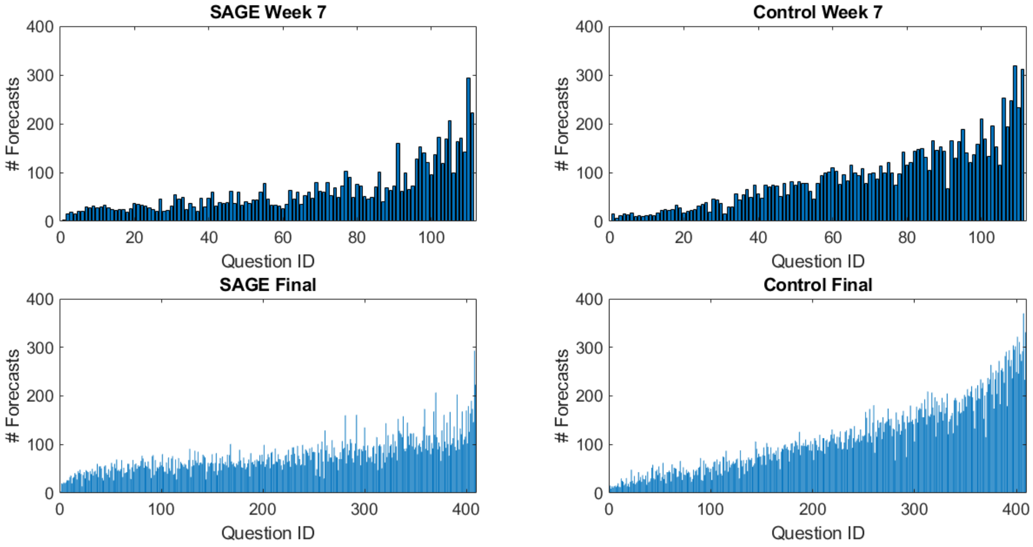


Figure 5.3: Question popularity for the SAGE and Control conditions at week 7 and at the end of the competition (Final). Questions are sorted by the total number of forecasts summed across both conditions.

whether the difference in Brier scores changed from before to after the intervention. Results from this test were ambiguous (-0.040 vs 0.026, t(397)=-2.09, BF_{10} = 0.99, Cauchy prior with scale = 0.707). As a result, it is unclear if limiting choice in this way confers a benefit to the aggregate.

Table 5.4: Human-only aggregate performance for the SAGE and control conditions depending on the degree of forecaster choice permitted to SAGE forecasters. During the first seven weeks of the competition, forecasters on the SAGE team were permitted to choose any questions that had not yet resolved (Full Choice). From the eighth week onward, forecasters were only permitted to choose from the 10 least popular questions when providing non-update forecasts (Limited Choice). Performance is measured with Mean of the Mean Daily Brier score (MMDB) and the Cohen's d of the SAGE aggregate relative to the Control aggregate.

Forecaster Choice	N	SAGE MMDB	Control MMDB	Cohen's d
Full	94	0.389	0.349	-0.129
Limited	295	0.319	0.345	0.090

5.4 Discussion

Incorporating algorithmic forecasts through hybridized aggregation conferred benefits over a human-only approach and allowed the SAGE team to produce the most accurate forecasts of any competitor. These benefits were most statistically reliable when the hybrid forecast averaged the human-only aggregate and algorithmic forecasts via M1. However, the absolute improvement in accuracy as measured by Brier score were largest with M2, which treated the algorithm as approximately equivalent to six human forecasters. This aggregate relied more heavily on the algorithmic forecasts when human forecasts were sparse, known to be outdated, or non-expert. The benefits from hybridization became more substantial when few human forecasts were made available for aggregation. This reflects the capacity of algorithmic forecasts to aide systems which need to scale to large numbers of questions.

Limiting forecaster choice to the least popular questions guaranteed a more even distribution of forecasters across questions. It is unclear if this restriction improved aggregate accuracy or if the average improvement observed from before to after the intervention was due to chance. If limiting choice improves aggregate accuracy, this would be surprising in light of previous research that had indicated that self-direction improves aggregate performance for crowds answering general knowledge questions [10], but may indicate that there are limits on the degree of self-direction that optimizes crowd accuracy.

This competition highlighted the potential for interdisciplinary collaboration between behavioral scientists and machine learning researchers. As a result of the joint efforts between these disciplines, our team outperformed what would have been possible by relying on either domain alone. More broadly, large-scale competitions like this provide an opportunity to test theories generated from laboratory studies in applied settings.

Chapter 6

General Discussion

In this dissertation, I argued for the theoretical importance of metacognition in crowd settings and demonstrated that letting crowd members choose which questions to answer improved aggregate accuracy for general knowledge questions. I jointly modeled the cognitive and metacognitive features of crowd members answering questions in an opt-in setting to explicate the relationship between choosing behavior and latent knowledge. I showed that impossible questions can be used to estimate metacognitive ability and that this metacognitive measure is a better predictor of a crowd member's contributions than the cognitive measure of accuracy. As a result, I was able to create highly accurate crowds with very few crowd members by selecting for metacognitive ability via impossible questions. I detailed my contributions to a geopolitical forecasting competition in which I developed models which aggregated human and algorithmic forecasts to produce the most accurate forecasts in the competition. Nonetheless, some aspects of the intersection between metacognition and aggregate performance present the opportunity for future research.

It would be valuable to identify which circumstances do not benefit from letting crowd members opt-in. It is likely that this occurs when participants exhibit certain metacognitive failures. While overconfidence is a common metacognitive failure, letting crowd members opt-in would likely remain beneficial so long as crowd members exhibit the *metacognitive sensitivity* [31] required to discriminate between their accurate and inaccurate responses based on confidence. A lack of metacognitive sensitivity is much rarer than overconfidence, although it is observed in certain populations such as those with radical beliefs [89]. It can also be induced via metacognitive inflation [86]. Future research can test this hypothesis and identify limits on the optimal degree of crowd member self-direction.

One challenge when testing the effect of an intervention on aggregate accuracy, such as the effect of letting crowd members opt-in, is a lack of statistical power. This lack of power stems from the fact that most comparisons between aggregates have statistical power that increases with the number of questions, and yet each of these questions requires many individuals to form a crowd. As a result, it is difficult to identify interventions that improve aggregate accuracy unless the effect, the number of questions, and the number of participants are very large. For instance, in Chapter 2 it was not possible to determine which of the "full choice" or "partial choice" conditions resulted in more accurate crowds. Even in the Hybrid Forecasting Competition, wherein 600 forecasters worked each week for nine months on hundreds of questions, it was unclear if limiting forecaster choice conferred a benefit to aggregate accuracy. Rather, the statistical test to evaluate this difference indicated that an observer should not change their relative credence in the null or alternative hypotheses (BF_{10} = 0.99) despite the substantial observed difference in aggregate accuracy. Cognitive models allow researchers to partially circumvent these statistical issues. In Chapter 3, for instance, I used a Bayesian cognitive model to estimate how aggregate accuracy of crowds that opt-in would vary depending on the difficulty of tasks or the abilities of individuals. More broadly, when a cognitive models provides a sufficiently accurate and detailed understanding of the cognitive and metacognitive properties of a task, researchers can estimate aggregate effects even when data is limited.

Much of the research contained in this dissertation has interdisciplinary origins. The studies reported in Chapter 2 were conducted by hypothesizing that the beneficial effects of self-direction in educational contexts (e.g. [106] [53]) might extended into an aggregation context. In the Hybrid Forecasting Competition the collaborative efforts of the behavioral scientists and machine learning researchers on the SAGE team allowed us to generate the most accurate forecasts in the competition. Taking a broad perspective, interdisciplinary work allows researchers to apply findings from one domain into another in order to discover novel insights.

Chapter 7

Estimating COVID-19 Antibody Seroprevalence in Santa Clara County, California. A re-analysis of Bendavid et al.

This section details an analysis of the prevalence of COVID-19 in Santa Clara using a simple Bayesian model. While this work has little direct relevance to the topics discussed in this dissertation, it is a valuable scientific contribution that may be of interest to readers.

> Bennett, S. T., & Steyvers, M. Estimating COVID-19 Antibody Seroprevalence in Santa Clara County, California. A re-analysis of Bendavid et al. *medRxiv*, 2020.

A recent study by Bendavid et al. claimed that the rate of infection of COVID-19 in Santa Clara county was between 2.49% and 4.16%, 50-85 times higher than the number of officially confirmed cases. The statistical methodology used in that study overestimates of rate of infection given the available data. We jointly estimate the sensitivity and specificity of the test kit along with rate of infection with a simple Bayesian model, arriving at lower estimates of the rate of COVID-19 in Santa Clara county. Re-analyzing their data, we find that the rate of infection was likely between

0.27% and 3.21%.

7.1 Introduction

A recent Stanford study by Bendavid et al. [5] found that 50 of 3330 people living in Santa Clara county tested positive for the novel Coronavirus, COVID-19. The authors of the study claim that this rate of positive test results, combined with the specificity and sensitivity of the test kit they used, were sufficient to believe that 2.49% to 4.16% of the people living in Santa Clara county were infected with COVID-19 as of April 4th, 2020. This study was quick to make headlines [27] [95] [105], but those estimates are too high given the data available.

The specificity of the test kit is too low to come to such confidently high estimates based on their sample. The authors give confidence intervals for the specificity of the test kit that ranged from 98.1% to 100%. If we take this range seriously, then it would be possible to explain their results even if no one in Santa Clara were sick. Indeed, a sample of 3330 healthy people tested by a test kit with 98.1% specificity would have 63.27 false positives on average, more than the positive test results found by Bendavid et al.

In this paper, we provide the results of a simple Bayesian model in which the specificity and sensitivity of the test kit are jointly inferred along with the rate of infection in Santa Clara at large. This model allows us to correctly propagate uncertainty about the accuracy of the test kit when estimating the rate of infection. Under our model, we infer the posterior density of the rate of infection, specificity, and sensitivity and compute 95% Highest Density Intervals for each of them to summarize our findings.

We focus solely on the statistical methodology used to infer the rate of infection from the sample and make no claims about the appropriateness of the other methods used in the Bendavid et al. study, such as demographic re-weighting or the sampling procedure.

7.2 Methods

7.2.1 Data

In order to inform the specificity and sensitivity of the test kit, we assume “Scenario 3” as described by the authors. That is, we assume that both their data and the data provided by the manufacturer are useful for estimating the specificity and sensitivity of the test kit used on their Santa Clara sample. When the manufacturer assessed their test kit, it was able to correctly identify 369 of 371 pre-COVID samples known to be uninfected. The authors tested an additional 30 pre-COVID samples from hip surgery, all of which the test kit correctly identified. Combining both sources, the test kit falsely identified 2 out of 401 samples that could not have had the disease. To inform its ability to detect the disease in infected samples, the manufacturer found the test kit correctly identified 153 of 160 samples from clinically confirmed COVID-19 patients. The authors tested samples from an additional 37 clinically confirmed patients, 25 of which were correctly identified. In total, the test kit correctly identified 178 of 197 presumed-infected individuals.

The authors collected 3330 volunteers from Santa Clara county recruited via Facebook and tested them with this test kit. 50 of these people tested positive for COVID-19.

7.2.2 Priors

We assume a-priori that sensitivity, specificity, and the rate of infection are each distributed according to a Beta(2,2) distribution. The probability density function associated of this distribution is in panel 1 of Figure 7.2. This prior is uninformative, reflecting the belief that these variables could take any of a wide range of possible values. Sensitivity to these priors is included in the Results section.

7.2.3 Model Specification

We assume a simple Bayesian model that explains the test results given the uncertain specificity (α), sensitivity (β), and rate of infection in Santa Clara (π). From these values, we compute the rate of positive test results (ρ). The Santa Clara sample (x_3) is distributed according to a Binomial distribution given ρ. The test kit data (x_1 & x_2) are distributed Binomially with rate parameters based on the sensitivity and specificity of the test kit. A simple graphical representation is shown in Figure 7.1. Circles that are shaded gray are our observed data. Circles that are not shaded we infer by sampling. This is done jointly for all variables and all data.

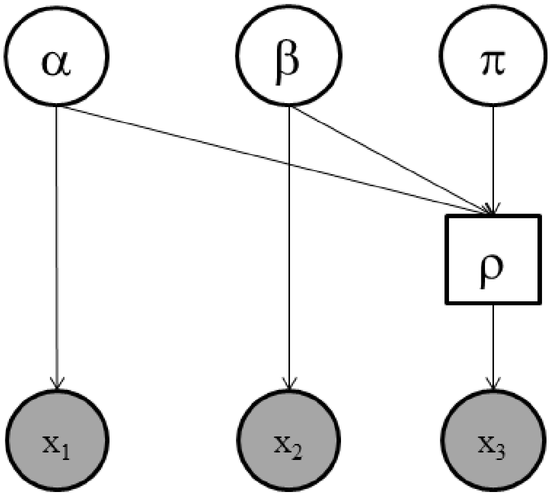


Figure 7.1: Graphical representation of our model.

7.2.4 Sampling Details and Code

We use JAGS, a Markov chain Monte Carlo (MCMC) sampler, to condition the model upon the data. We collected 5,000,000 samples of the posterior distribution from 7 chains after 1,000,000 burn-in samples.

All data and code associated with the project is available via the Open Science Framework: https://osf.io/5qhgb/.

7.2.5 Demographic re-weighting

Bendavid et al. re-weight the estimate of the rate of infection based on the demographic data of their sample. This factor was intended to compensate for the non-representative sample they recruited for their study. They propagate uncertainty from the inferred number of actual cases in their sample to the inferred number of actual cases in the population at large by multiplying the sample's rate of infection by 1.87. We make no claim about the suitability of this re-weighting, but include it in order to match our results with those of the original authors. Results are presented both with and without this factor.

7.3 Results

7.3.1 Specificity and sensitivity

The posterior distributions for specificity and sensitivity are shown in panels 3 and 4 of figure 7.2. There is substantial uncertainty about both of these values based on the data available. The 95% highest density interval for the specificity of the test kit is [98.73%, 99.83%] and the 95% highest density interval for the sensitivity of the test kit is [85.11%, 93.56%]. If this lower bound of specificity (98.72%) were true, in expectation we would observe 42.6 positive test results out of the 3330 individuals tested even if no one in Santa Clara were infected with COVID-19.

7.3.2 Rate of Infection

The posterior distribution of the rate of sick individuals in Santa Clara county is shown in panel 4 of Figure 7.2. The 95% highest density interval associated with this posterior distribution is [0.27%, 1.72%]. Replicating the demographic re-weighting done by the original authors, this

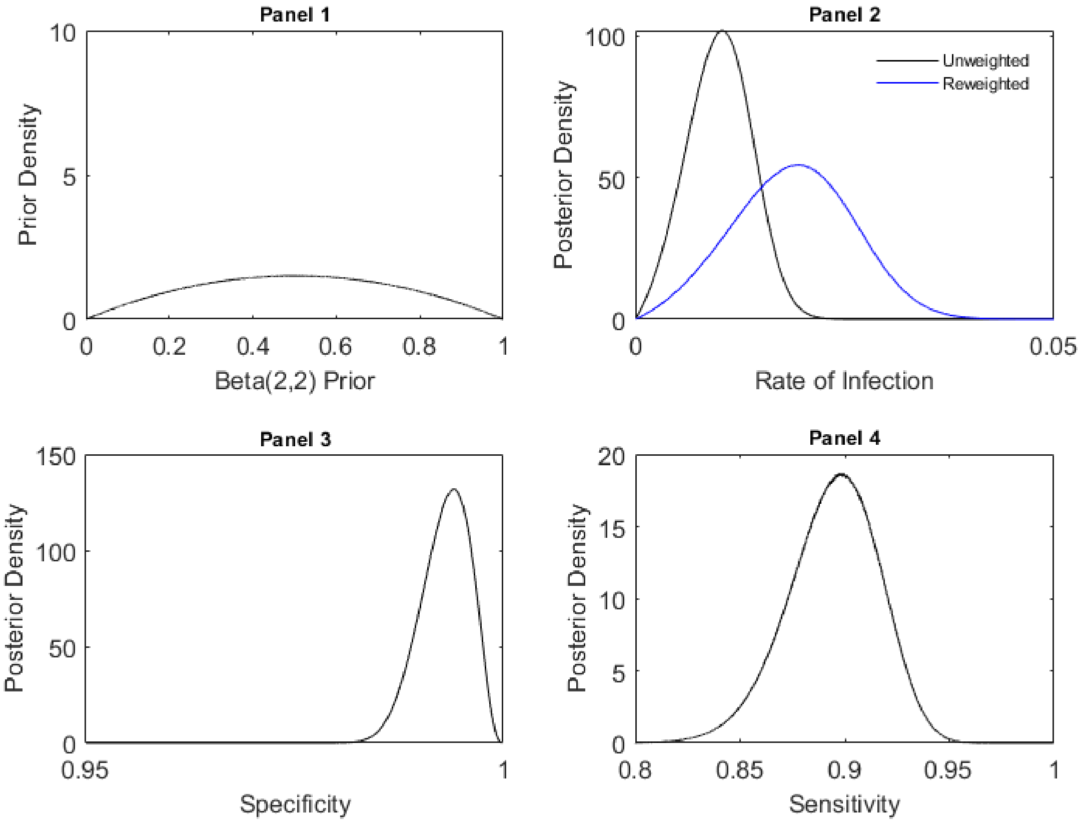


Figure 7.2: **Panel 1**: Prior distribution for Specificity, Sensitivity, and the Rate of Infection. **Panel 2**: Posterior distribution of the rate of infection in Santa Clara, before and after re-weighting based on demographic data. **Panel 3**: Posterior distribution of the specificity of the test kit. **Panel 4**: Posterior distribution of the sensitivity of the test kit.

interval becomes [0.49%, 3.21%].

7.3.3 Prior Sensitivity Analysis

We analyzed how our results differed as a function of the priors used in the model. Table 7.1 shows the results assuming all distributions have prior Beta(x,x) for various values of x. These results are each based on 500,000 samples from JAGS after 100,000 burn-in samples.

x	Specificity 95HDI	Sensitivity 95HDI	Infection 95HDI
1	[98.66% 99.89%]	[85.59% 93.92%]	[0.01% 3.18%]
2	[98.72% 99.82%]	[85.12% 93.53%]	[0.56% 3.44%]
4	[98.74% 99.72%]	[83.99% 92.80%]	[0.8% 3.22%]
8	[98.66% 99.56%]	[82.12% 91.30%]	[1.18% 3.41%]
16	[98.46% 99.32%]	[78.54% 88.47%]	[1.79% 4.01%]

Table 7.1: Sensitivity of our results based on our priors. 95% Highest density intervals for the variables of interest are presented for several possible priors.

7.4 Discussion

Reanalyzing the data from Bendavid et al., we find that the rate of infection was likely between 0.27% and 3.21% in early April. This interval is substantially lower than the interval used to draw conclusions by the original authors (2.49% to 4.16%). Based on our analyses, there is a 79.47% chance that the true rate of infection was below the lower bound of this interval even after re-weighting based on demographics.

We re-analyze some of the conclusions of the original authors based on our estimate of the rate of infection, 0.27% to 3.21%. This rate of infection means between 5,000 and 65,000 people were infected in Santa Clara county. As of April 1st, 956 cases had been confirmed in Santa Clara. This would correspond to between a 5-to-1 and 65-to-1 ratio of observed-to-total cases of COVID-19 (the underascertainment rate).

Bibliography

[1] P. Atanasov, P. Rescober, E. Stone, S. A. Swift, E. Servan-Schreiber, P. Tetlock, L. Ungar, and B. Mellers. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management science*, 63(3):691–706, 2017.

[2] S. Atir, E. Rosenzweig, and D. Dunning. When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, 26(8):1295–1303, 2015.

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

[4] A. Beger and M. D. Ward. Assessing amazon turker and automated machine forecasts in the hybrid forecasting competition. In *7th Annual Asian Political Methodology Conference*, pages 5–6, 2019.

[5] E. Bendavid, B. Mulaney, N. Sood, S. Shah, E. Ling, R. Bromley-Dulfano, C. Lai, Z. Weissberg, R. Saavedra, J. Tedrow, D. Tversky, A. Bogan, T. Kupiec, D. Eichner, R. Gupta, J. Ioannidis, and J. Bhattacharya. Covid-19 antibody seroprevalence in santa clara county, california. *medRxiv*, 2020.

[6] A. Benjamin and B. H. Ross. *The Psychology of Learning and Motivation: Skill and Strategy in Memory Use*, volume 48. Academic Press, 2008.

[7] A. S. Benjamin. Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. *Psychology of learning and motivation*, 48:175–223, 2008.

[8] A. S. Benjamin and R. D. Bird. Metacognitive control of the spacing of study repetitions. *Journal of Memory and Language*, 55(1):126–137, 2006.

[9] A. S. Benjamin and B. H. Ross. *The psychology of learning and motivation: skill and strategy in memory use*, volume 48. Academic Press, 2008.

[10] S. T. Bennett, A. S. Benjamin, P. K. Mistry, and M. Steyvers. Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. *Computational Brain & Behavior*, 1(1):90–99, 2018.

[11] D. Bensch, D. L. Paulhus, L. Stankov, and M. Ziegler. Teasing apart overclaiming, overconfidence, and socially desirable responding. *Assessment*, 26(3):351–363, 2019.

[12] Y. Bereby-Meyer, J. Meyer, and D. V. Budescu. Decision making under internal uncertainty: The case of multiple-choice tests with different scoring rules. *Acta psychologica*, 112(2):207–220, 2003.

[13] M. Boekaerts and A. Minnaert. Self-regulation with respect to informal learning. *International journal of educational research*, 31(6):533–544, 1999.

[14] E. Bokhari and L. Hubert. The lack of cross-validation can lead to inflated results and spurious conclusions: A re-analysis of the macarthur violence risk assessment study. *Journal of Classification*, 35(1):147–171, 2018.

[15] J. V. Bradley. Overconfidence in ignorant experts. *Bulletin of the Psychonomic Society*, 17(2):82–84, 1981.

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

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

[18] D. V. Budescu and E. Chen. Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2):267–280, 2015.

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

[20] S. Castano, A. Ferrara, and S. Montanelli. Leveraging crowd skills and consensus for collaborative web-resource labeling. *Future Generation Computer Systems*, 95:790–801, 2019.

[21] E. Chen, D. V. Budescu, S. K. Lakshmikanth, B. A. Mellers, and P. E. Tetlock. Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis*, 13(2):128–152, 2016.

[22] S. A. Culpepper and J. J. Balamuta. A hierarchical model for accuracy and choice on standardized tests. *Psychometrika*, pages 1–26, 2015.

[23] C. P. Davis-Stober, D. V. Budescu, J. Dana, and S. B. Broomell. When is a crowd wise? *Decision*, 1(2):79, 2014.

[24] P. D. Dunlop, J. S. Bourdage, R. E. De Vries, I. M. McNeill, K. Jorritsma, M. Orchard, T. Austen, T. Baines, and W.-K. Choe. Liar! liar!(when stakes are higher): Understanding how the overclaiming technique can be used to measure faking in personnel selection. *Journal of Applied Psychology*, 2019.

[25] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015.

[26] C. Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 162–170, 2018.

[27] C. Farr. Antibody study suggests covid-19 could be far more prevalent in the bay area than official numbers suggest. `https://www.cnbc.com/2020/04/17/santa-clara-covid-19-antibody-study-suggests-broad-asymptomatic-spread.html`, 2020. Accessed: 2020-04-22.

[28] J. L. Fiechter, A. S. Benjamin, and N. Unsworth. 16 the metacognitive foundations of effective remembering. *The Oxford Handbook of Metamemory*, page 307, 2016.

[29] J. R. Finley, J. G. Tullis, and A. S. Benjamin. Metacognitive control of learning and remembering. In *New Science of Learning*, pages 109–131. Springer, 2010.

[30] S. M. Fleming and N. D. Daw. Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, 124(1):91, 2017.

[31] S. M. Fleming and H. C. Lau. How to measure metacognition. *Frontiers in human neuroscience*, 8:443, 2014.

[32] S. Frederick. Cognitive reflection and decision making. *Journal of Economic perspectives*, 19(4):25–42, 2005.

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

[34] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

[35] D. J. Gerner, P. A. Schrodt, O. Yilmaz, and R. Abu-Jabr. The creation of cameo (conflict and mediation event observations): An event data framework for a post cold war world. In *annual meeting of the American Political Science Association*, volume 29, 2002.

[36] M. Goldsmith and A. Koriat. The strategic regulation of memory accuracy and informativeness. *Psychology of learning and motivation*, 48:1–60, 2007.

[37] Y. Grushka-Cockayne, V. R. R. Jose, and K. C. Lichtendahl Jr. Ensembles of overfit and overconfident forecasts. *Management Science*, 63(4):1110–1130, 2017.

[38] E. Heim, T. Roß, A. Seitel, K. März, B. Stieltjes, M. Eisenmann, J. Lebert, J. Metzger, G. Sommer, A. W. Sauter, et al. Large-scale medical image annotation with crowd-powered algorithms. *Journal of Medical Imaging*, 5(3):034002, 2018.

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

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

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

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

[43] JASP Team. JASP (Version 0.12)[Computer software], 2020.

[44] H. Jeffreys. *Theory of Probability*. Oxford, 1961.

[45] H. Jeffreys. *Theory of probability, 3rd edn oxford: Oxford university press*. 1961.

[46] B. I. Johanna and M. van der Heijden. The development and psychometric evaluation of a multidimensional measurement instrument of professional expertise. *High Ability Studies*, 11(1):9–39, 2000.

[47] M. Z. Juni and M. P. Eckstein. The wisdom of crowds for visual search. *Proceedings of the National Academy of Sciences*, 114(21):E4306–E4315, 2017.

[48] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[49] C. M. Kelley and L. Sahakyan. Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language*, 48(4), 2003.

[50] A. Kepecs and Z. F. Mainen. 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, 2012.

[51] V. Klusmann, A. Evers, R. Schwarzer, and I. Heuser. A brief questionnaire on metacognition: psychometric properties. *Aging & mental health*, 15(8):1052–1062, 2011.

[52] A. Koriat and M. Goldsmith. Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological review*, 103(3):490, 1996.

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

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

[55] D. A. Lagnado and S. Sloman. The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(4):856, 2004.

[56] M. D. Lee and I. Danileiko. Using cognitive models to combine probability estimates. *Judgment and Decision Making*, 9(3):259, 2014.

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

[58] Q. Li and P. K. Varshney. Does confidence reporting from the crowd benefit crowdsourcing performance? In *Proceedings of the 2nd International Workshop on Social Sensing*, pages 49–54, 2017.

[59] F. Lieder and T. L. Griffiths. When to use which heuristic: A rational solution to the strategy selection problem. In *CogSci*, 2015.

[60] F. Lieder, D. Plunkett, J. B. Hamrick, S. J. Russell, N. Hay, and T. Griffiths. Algorithm selection by rational metareasoning as a model of human strategy selection. In *Advances in Neural Information Processing Systems*, pages 2870–2878, 2014.

[61] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

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

[63] B. Maniscalco. Type 2 signal detection theory analysis using meta-d. `http://www.columbia.edu/~bsm2105/type2sdt/`, 2020. Accessed: 2020-07-27.

[64] B. Maniscalco and H. Lau. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, 21(1):422–430, 2012.

[65] B. Maniscalco and H. Lau. Signal detection theory analysis of type 1 and type 2 data: meta-d, response-specific meta-d, and the unequal variance sdt model. In *The cognitive neuroscience of metacognition*, pages 25–66. Springer, 2014.

[66] A. E. Mannes, J. B. Soll, and R. P. Larrick. The wisdom of select crowds. *Journal of personality and social psychology*, 107(2):276, 2014.

[67] K. A. Martire, B. Growns, and D. J. Navarro. What do the experts know? calibration, precision, and the wisdom of crowds among forensic handwriting experts. *Psychonomic bulletin & review*, 25(6):2346–2355, 2018.

[68] B. Mellers, E. Stone, P. Atanasov, N. Rohrbaugh, S. E. Metz, L. Ungar, M. M. Bishop, M. Horowitz, E. Merkle, and P. Tetlock. The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of experimental psychology: applied*, 21(1):1, 2015.

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

[70] E. C. Merkle, M. Steyvers, B. Mellers, and P. Tetlock. A neglected dimension of good forecasting judgment: The questions we choose also matter. *International Journal of Forecasting*, in press.

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

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

[73] J. Mezirow. A critical theory of adult learning and education. *Adult education quarterly*, 32(1):3–24, 1981.

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

[75] F. Morstatter, A. Galstyan, G. Satyukov, D. Benjamin, A. Abeliuk, M. Mirtaheri, K. Hossain, P. Szekely, E. Ferrara, A. Matsui, et al. Sage: a hybrid geopolitical event forecasting system. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6557–6559. AAAI Press, 2019.

[76] A. E. Murr. "wisdom of crowds"? a decentralised election forecasting model that uses citizens' local expectations. *Electoral Studies*, 30(4):771–783, 2011.

[77] T. O. Nelson and L. Narens. The psychology of learning and motivation. *Metamemory: A theoretical framework and new findings*, 1990.

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

[79] S. Oyama, Y. Baba, Y. Sakurai, and H. Kashima. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *Twenty-third International Joint Conference on Artificial Intelligence*, 2013.

[80] S. G. Paris and A. H. Paris. Classroom applications of research on self-regulated learning. *Educational psychologist*, 36(2):89–101, 2001.

[81] T. Peeters. Testing the wisdom of crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting*, 34(1):17–29, 2018.

[82] G. Pennycook and D. G. Rand. Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality*, 88(2):185–200, 2020.

[83] L. Pion-Tonachini, S. Makeig, and K. Kreutz-Delgado. Crowd labeling latent dirichlet allocation. *Knowledge and information systems*, 53(3):749–765, 2017.

[84] D. Prelec, H. S. Seung, and J. McCoy. Finding truth even if the crowd is wrong. In *Working paper, MIT*. 2013.

[85] D. Prelec, H. S. Seung, and J. McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541(7638):532–535, 2017.

[86] D. Rahnev and S. M. Fleming. How experimental procedures influence estimates of metacognitive ability. *Neuroscience of consciousness*, 2019(1):niz009, 2019.

[87] R. B. Rao, G. Fung, and R. Rosales. On the dangers of cross-validation. an experimental evaluation. In *Proceedings of the 2008 SIAM international conference on data mining*, pages 588–596. SIAM, 2008.

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

[89] M. Rollwage, R. J. Dolan, and S. M. Fleming. Metacognitive failure as a feature of those holding radical beliefs. *Current Biology*, 28(24):4014–4021, 2018.

[90] J. N. Rouder. Bayes factor calculators. `http://pcl.missouri.edu/bayesfactor`, 2014. Accessed: 2020-07-27.

[91] J. N. Rouder. Bayes factor calculator. *Website. Retrieved from http://pcl.missouri.edu/bayesfactor*, 2018. Accessed: 2018-04-23.

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

[93] G. Rowe and G. Wright. The delphi technique as a forecasting tool: issues and analysis. *International journal of forecasting*, 15(4):353–375, 1999.

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

[95] K. Selig. Santa clara county covid-19 cases could be 50 to 85 times higher than reported, stanford study finds. `https://www.stanforddaily.com/2020/04/17/santa-clara-county-covid-19-cases-could-be-50-to-85-times-higher-than-reported-sta` 2020. Accessed: 2020-04-22.

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

[97] D. M. Sobel and T. Kushnir. The importance of decision making in causal learning from interventions. *Memory & Cognition*, 34(2):411–419, 2006.

[98] B. J. Stastny and P. E. Lehner. Comparative evaluation of the forecast accuracy of analysis reports and a prediction market. *Judgment & Decision Making*, 13(2), 2018.

[99] M. Steyvers and B. Miller. Cognition and collective intelligence. In M. Bernstein and T. W. Malone, editors, *Handbook of Collective Intelligence*, pages 119–138. MIT Press, 2015.

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

[101] M. Steyvers, J. B. Tenenbaum, E.-J. Wagenmakers, and B. Blum. Inferring causal networks from observations and interventions. *Cognitive science*, 27(3):453–489, 2003.

[102] J. Sun, S. Moosavi, R. Ramnath, and S. Parthasarathy. Qdee: question difficulty and expertise estimation in community question answering sites. *Twelfth International AAAI Conference on Web and Social Media*, 2018.

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

[104] P. E. Tetlock, B. A. Mellers, N. Rohrbaugh, and E. Chen. Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science*, 23(4):290–295, 2014.

[105] L. A. Times. Stanford antibody study estimates covid-19 infected 50 to 85 times more people than testing identified in santa clara county. `https://ktla.com/news/california/stanford-antibody-study-estimates-covid-19-infected-at-least-50-times-more-people-t` 2020. Accessed: 2020-04-22.

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

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

[108] L. Ungar, B. Mellers, V. Satopää, P. Tetlock, and J. Baron. The good judgment project: A large scale test of different methods of combining expert predictions. In *2012 AAAI Fall Symposium Series*, 2012.

[109] E.-J. Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, 14(5):779–804, 2007.

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

[111] P. Welinder, S. Branson, P. Perona, and S. J. Belongie. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, pages 2424–2432, 2010.

[112] R. Williams et al. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *Stata Journal*, 6(1):58, 2006.

[113] P. H. Winne and A. F. Hadwin. Studying as self-regulated learning. *Metacognition in educational theory and practice*, 93:27–30, 1998.

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

[115] J. Wolfers and E. Zitzewitz. Prediction markets. *Journal of economic perspectives*, 18(2):107–126, 2004.

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

[117] B. J. Zimmerman. A social cognitive view of self-regulated academic learning. *Journal of educational psychology*, 81(3):329, 1989.

Appendix A

Appendix for Making a wiser crowd: Benefits of individual metacognitive control on crowd performance

Distribution of responses in the hard condition of Experiment 1 and the partial opt-in and control conditions in Experiment 2 are shown below.

Figure A.1: Question responses for the self-directed and control participants in the hard condition of Experiment 1. Questions are 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.

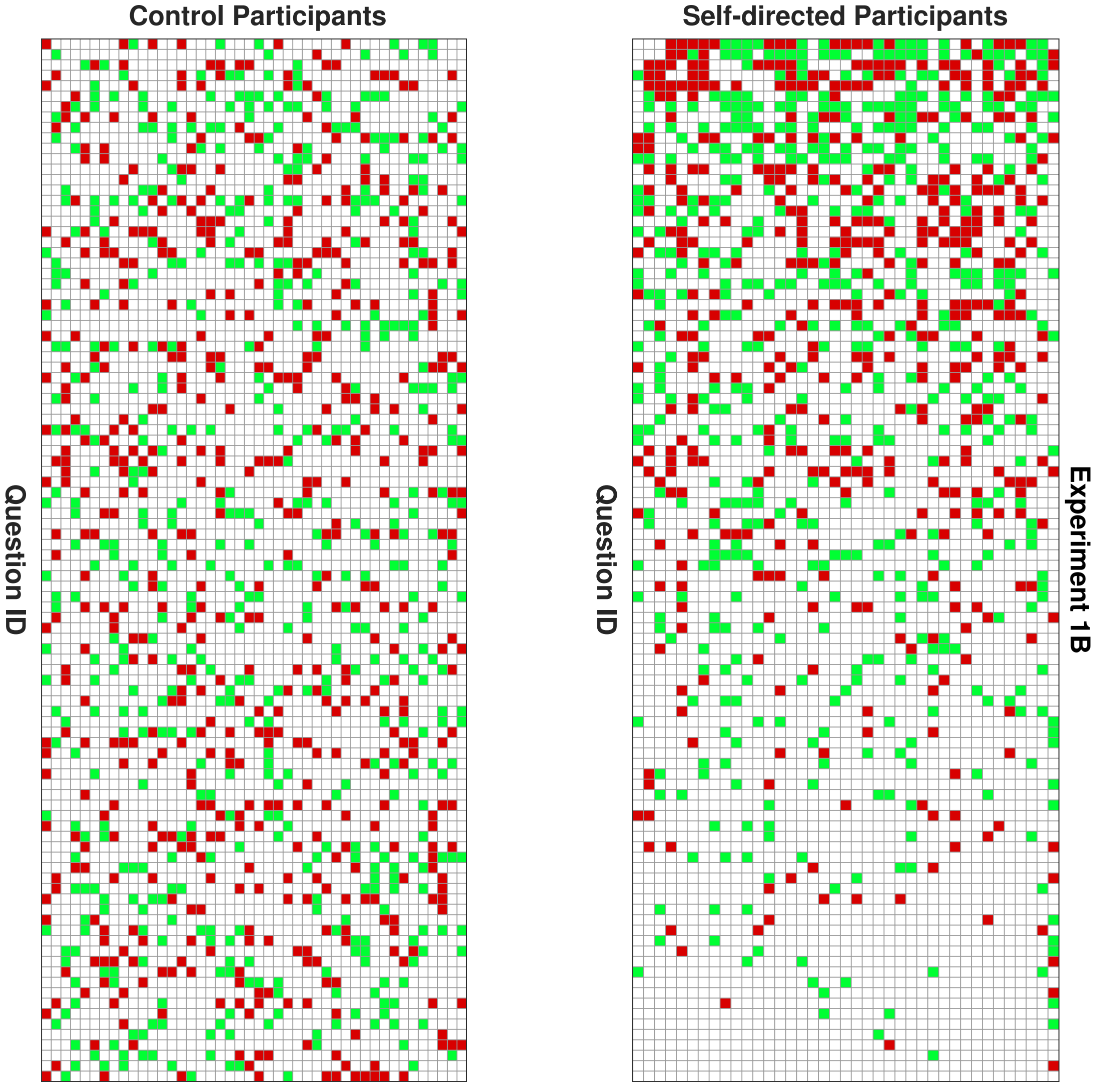
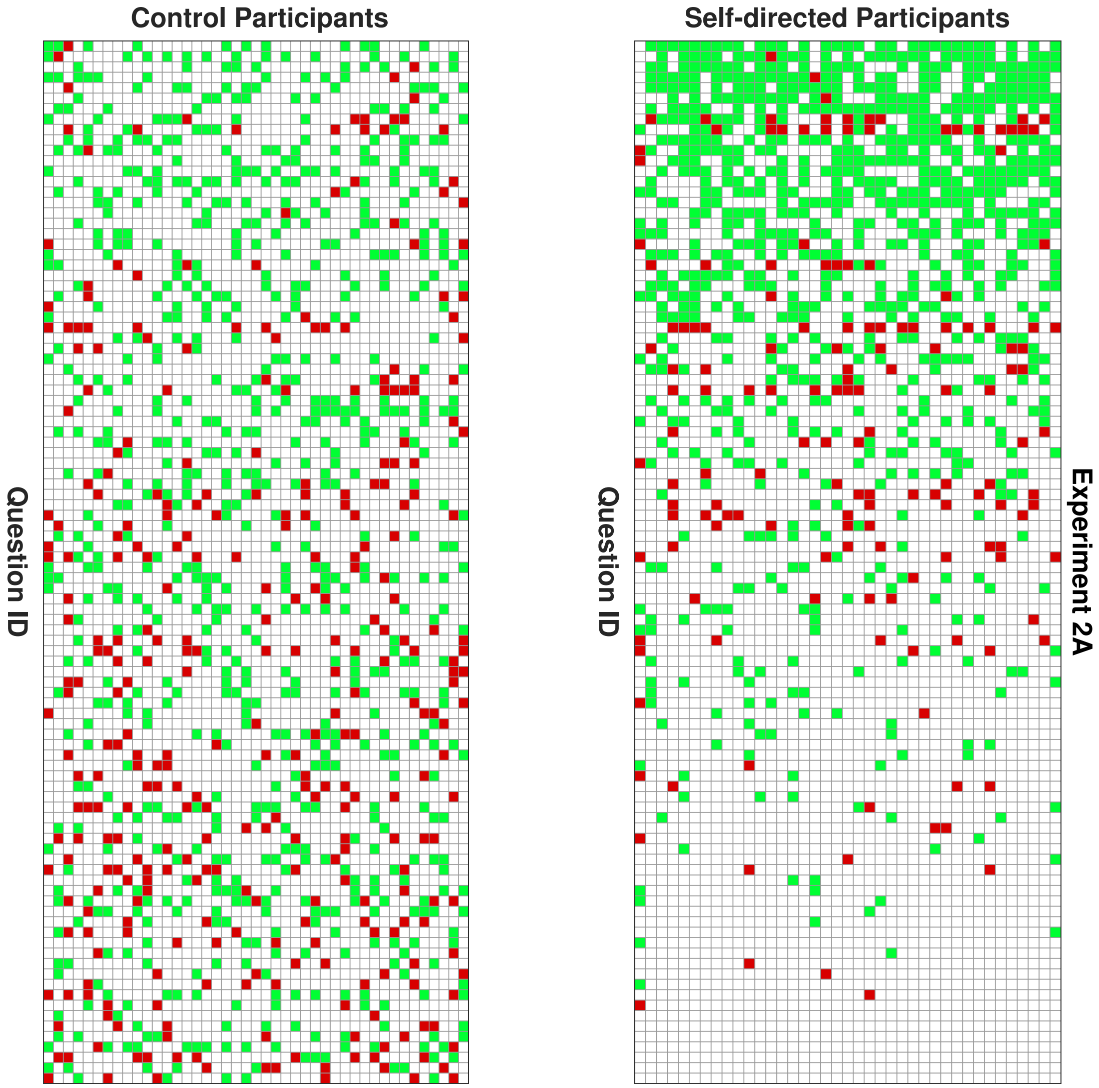


Figure A.2: Question responses from the partial opt-in and control conditions of Experiment 2. Questions are sorted by the number of participants who selected the question in the self-directed condition. Green squares represent correct responses, red squares represent incorrect responses, and white squares represent no response.



Appendix B

Appendix for Leveraging metacognitive ability to improve crowd accuracy via impossible questions.

Confidence ratings can be used to measure metacognitive ability, for instance by computing *meta-d'* or measuring participant calibration. However, in an experimental context where participants opt-in, participants may avoid giving any response when they lack confidence. In this way, opting-in would act as a filter that prevents us from observing confidence ratings that would normally be associated with a metacognitive judgment of doubt. Indeed, in a crowdsourcing context in which participants could opt-in, there was no benefit to crowd performance when leveraging confidence ratings [58]. This finding would be unsurprising if there were little additional metacognitive signal in confidence ratings after accounting for opt-in behavior. Nonetheless, we compared IQC (an individual's propensity to skip impossible questions) to these other metacognitive measures.

We computed meta-d', the ratio between meta-d' and d', and overconfidence. We computed d' and meta-d' with publicly available software [64, 65, 63]. We converted the scalar ratings solicited in our experiment into categorical ratings in order to compute meta-d'. To do this, we treated all confidence ratings above the median confidence rating of all participants (93.7%) as high confidence and those below the median as low confidence. IQC was positively correlated with the measure of participant expertise, d' (r = 0.50, BF = 16.3). However, the relationship between IQC and the metacognitive measures is less clear: the observed correlations between IQC and both meta-d' (r = 0.24, BF = 0.72) and $\frac{meta-d'}{d'}$ (r = -0.11, BF = 0.39) are likely spurious. In addition to meta-d', we evaluated overconfidence by computing the difference between the average accuracy of each participant and their average confidence. Participants were overconfident to the degree that their average confidence ratings were higher than their average accuracy. IQC was negatively correlated with overconfidence, although this may have been due to chance as well (r = -0.30, BF = 1.12). These ambiguous findings may be due to the fact that low confidence responses are censored in an opt-in context. Future research may be able to more clearly establish what relationship exists between IQC and other metacognitive measures.