AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

© 2022 American Psychological Association
ISSN: 2325-9965

2022, Vol. 9, No. 1, 60–73
https://doi.org/10.1037/dec0000165

Leveraging Metacognitive Ability to Improve Crowd Accuracy via Impossible Questions

Stephen T. Bennett and Mark Steyvers

Department of Cognitive Sciences, University of California, Irvine

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 measure expertise with the participant's tendency to skip impossible questions. These questions have no correct answers and serve as a metacognitive measure of a participant's ability to recognize when they lack knowledge. In contexts where individuals choose which questions to answer, those who are selective about when to contribute to the crowd are valuable. 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.

Keywords: wisdom of crowds, metacognition, expertise, impossible questions, overclaiming

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 (Atanasov et al., 2017; Beger & Ward, 2019; Turner et al., 2014) or prediction markets (Stastny & Lehner, 2018; Wolfers & Zitzewitz, 2004). Crowds are used to generate accurate labels for images (Welinder et al., 2010), electroencephalogram (EEG) components (Pion-Tonachini et al., 2017), music (Castano et al., 2019), and medical segmentations (Heim et al., 2018). While geopolitical forecasting and labeling are common applications, crowds are effective for a surprising breadth of tasks, including solving combinatorial problems (Yi et al., 2012), predicting the outcome of sporting events (Herzog & Hertwig, 2011; Peeters, 2018), identifying authorship from handwriting (Martire et al., 2018), and visual search (Juni & Eckstein, 2017). Despite the impressive history of crowds, aggregating across a large and diverse group does not guarantee accuracy (Davis-Stober et al., 2014; Burnap et al., 2015; Grushka-Cockayne et al., 2017).

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 (Mannes et al., 2014), contribution weighted scores (Chen et al., 2016), or genetic algorithms (Hill & Ready-Campbell, 2011). If historical accuracy is difficult or costly to obtain, questionnaires such as the cognitive reflection task (Frederick, 2005) allow the experimenter to select for more deliberative reasoners, which can be used to improve crowd accuracy (Eickhoff, 2018; Mellers et al., 2015). The cognitive reflection task can be thought of as an example of a set of seed questions—questions for which the

Stephen T. Bennett https://orcid.org/0000-0003-2756-9742

Stephen T. Bennett played a lead role in data curation, formal analysis, investigation, project administration, validation, visualization, writing—original draft and writing—review and editing, and an equal role in conceptualization and methodology. Mark Steyvers played a lead role in funding acquisition, supporting role in formal analysis, visualization, writing—original draft and writing—review and editing, and an equal role in conceptualization and methodology.

Correspondence concerning this article should be addressed to Stephen T. Bennett, Department of Cognitive Sciences, University of California, 2201 Social & Behavioral Sciences Gateway, Irvine, CA 92697-5100, United States. Email: stbennet@uci.edu

experimenter knows the answer—which are then used to identify the relative expertise of respondents on target questions of interest (Quigley et al., 2018). In a similar vein, instructional manipulation checks can identify which participants are attentive (Hauser & Schwarz, 2016) but may influence participant responses in detrimental ways (Hauser et al., 2018).

Another approach to improve crowd accuracy is to permit crowd members to select which questions to answer, or *opt-in* (Bennett et al., 2018). 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 (Merkle et al., 2017), indicating that metacognitive ability is related to forecasting ability. More accurate crowd members also tend to provide more coherent responses across items (Fan et al., 2019), which could result from a general metacognitive process. While not metacognition in the traditional sense of reasoning about one’s own reasoning, respondents’ expectations about the distribution of others’ judgments can be used to identify relative experts within a crowd (Prelec et al., 2013, 2017). Indeed, metacognition and domain skill are correlated (Johanna & van der Heijden, 2000), and so we should expect that participants’ metacognitive abilities relate to the accuracy of the resulting crowds. To measure metacognitive ability, psychometric methods typically require involved questionnaires (Klusmann et al., 2011). *Meta-d’* is another method that can differentiate between competency and metacognitive ability using a signal detection framework (Maniscalco & Lau, 2012), but again requires a large number of responses to questions with known ground truth. 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 nonexistent 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 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 metacognitive error on the part of the participant. The decision to answer or skip these questions could, however, be motivated by psychological factors other than a participant’s metacognitive ability. A participant’s propensity for risk-taking may lead them to answer questions even when they do not know the answer (Alnabhan, 2002; Campbell et al., 2004), participants may satisfice by skipping questions to reduce the effort required to complete a task (Bogner & Landrock, 2016; Krosnick et al., 1996), or otherwise may overstate their familiarity with topics due to impression management and positivity bias (Bensch et al., 2019; Bishop et al., 1980, 1986). Nonetheless, a participant’s decision to skip impossible questions, in particular, reflects their ability to identify when they lack knowledge and thereby demonstrates metacognitive skill.

A number of previous studies have utilized impossible questions. Bereby-Meyer et al. (2003) 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 (Atir et al., 2015). Overclaiming can be used to readily identify when people overstate their own knowledge and abilities (Bensch et al., 2019; Dunlop et al., 2020; Pennycook & Rand, 2020). Still, others term them nonsense questions and relate them to risk-taking (Alnabhan, 2002). Note that impossible questions require that the participant be allowed to opt-out of questions in the full experiment. If opting-out were only allowed for the impossible questions while all other questions required responses, then participants would be able to differentiate between the impossible questions and regular questions on that basis alone.

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 an experiment and reanalyze an existing data set by Bereby-Meyer et al. (2003) to examine how impossible questions can measure metacognitive ability and be used to improve crowd accuracy.

Experiment 1

Method

Participants

Thirty-five 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 Human Intelligence Task (HIT) approval rating on at least 1,000 HITs and had not participated in any of our previous studies that used overlapping questions. Each participant was compensated $5 for the 30 min the experiment was expected to take.

An immediate replication with 32 participants was conducted to assess the statistical reliability of the findings. While ordinarily analyses for separate waves of recruitment would be presented separately, this article will rely on Bayes factors (BFs) and credible intervals to determine how the data alter our beliefs in various hypotheses. One reason to avoid combining recruitment waves when using traditional statistical analysis is that the decision to include additional participants would dramatically alter the test statistics and *p*-values associated with our analyses (Kruschke & Liddell, 2018), which is an issue that Bayesian methods avoid. Additionally, we will not establish any cutoff between "significant" and "nonsignificant" findings that could lead to issues related to optional stopping.

Stimuli and data are available via the Open Science Framework: https://osf.io/8r3t9/?view_only=c5a9694d4900431ab00566e124a10b1d.

Stimuli

Stimuli consisted of 94 general knowledge binary-choice questions with a known ground truth from Bennett et al. (2018). The questions were drawn from 12 general topics: World Facts, World History, Sports, Earth Sciences, Physical Sciences, Life Sciences, Psychology, Space and Universe, Math and Logic, Climate Change, Physical Geography, and Vocabulary. Based on a previous experiment, this set of questions resulted in an average accuracy of 76%. In addition to the general knowledge questions, participants were asked six 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 1. The four impossible questions not included in that table are as follows: "When did the battle of Kavkav take place? (a) Before 1647 (b) After 1647," "How many sides does a Detseroid have? (a) More than 20 (b) Less than 20," "What is the population of Synomle? (a) More than 10,000 (b) Less than 10,000," and "What is the Parlichev method? (a) A way of extracting minerals from soil (b) A way of determining the distance from a star."

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

Table 1
Example General Knowledge and Impossible Questions

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?

Note. Correct answers are italicized. Impossible questions have no correct answers.

required to report how confident they were in their answer, from 50% to 100%[1].

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). Once completed, participants also gave confidence ratings for each of the 12 categories of questions, although these data will not be used in any analyses. Finally, participants were asked if they had any feedback for us, given the option to receive detailed feedback for all 100 questions, and thanked for their time.

Aggregating and Scoring Crowds

The primary outcome of interest is the relative performance of crowds composed of different individuals; which combinations of individuals result in the most competent crowds? To answer this question, there needs to be a common method of aggregating crowd members and scoring the resultant aggregates. Many such methods exist, and we will consider one common aggregation method scored with two common scoring rules to focus on the impact of including and excluding individuals when creating the crowd.

We will aggregate responses with *average confidence* (Ariely et al., 2000). Under this aggregation strategy, the probability assigned to a response option is equal to the average of all confidence ratings associated with that response option. Since all questions are binary forced choice, we assume that a participant who endorses the first response option with probability *X*% endorses the second response option with probability (100 − *X*)%[2]. Aggregating responses in this way captures the average endorsement of each response option weighted by crowd members' confidence. Participants who opt-out of a question are ignored when computing this aggregate.

As an example of aggregating with this method, suppose four individuals were asked the binary question "House flies have an average life span of less than 2 days, True or False?." One participant opts out, one places 75% confidence in the incorrect answer ("True"), and the others place 65% and 90% confidence in the correct answer ("False"). Under average confidence, we assign a 60% probability to "False" [i.e., (0.25 + 0.65 + 0.90)/3 = 0.6].

We will measure crowd performance with two methods: *linear loss* and *Brier score*. The linear loss for a single question is the difference between the probability assigned to the correct response option and 100%. The *Brier score* is computed by taking the square of this linear loss[3]. Linear loss provides an easily interpretable metric that matches intuitions associated with accuracy while the Brier score has the advantage of being a proper scoring rule and is frequently used in forecasting contexts (Witkowski et al., 2017). Returning to the hypothetical crowd above that answered a question about fly lifespans, the aggregate assigns 60% probability to the correct answer and therefore receives a linear loss of 0.4 and a Brier score of 0.16 (i.e., 1.0 − 0.6 = 0.4 and $0.4^2 = 0.16$).

We also aggregated crowds with Majority Rule, but those analyses are reported in Appendix A since the aggregates resulting from Majority Rule do not produce probability estimates. Those probability estimates are required to apply the linear loss and Brier scores.

Bayes Factors

For all analyses, we utilize 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 (Jarosz & Wiley, 2014; Kass & Raftery, 1995; Rouder et al., 2009; Wagenmakers, 2007), including the ability to detect evidence in favor of a null hypothesis and a straightforward interpretation. In order to compute the BFs, we used the software package JASP (JASP Team, 2020)

[1] Participants who opted out were also allowed to give a confidence rating but were not required to do so and it is unclear what that confidence corresponds to.

[2] In this way, we are assuming that participants satisfy the principle of countable additivity. Note that this assumption is unlikely to be true in practice (see, e.g., Mandel, 2008).

[3] We use the formulation of the Brier score that takes values from 0 to 1, which is computed by the formula BS = $(1 - c)^2$ where BS is the Brier score and *c* is the probability (from 0 to 1) assigned to the correct response option.

and a BF calculator available online (Rouder, 2014; Rouder et al., 2009). In both cases, we maintained the default priors that came with the software when analyzing data.

To interpret these BFs, 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. In order to improve readability, BFs larger than 100,000 are reported as BF > 100,000.

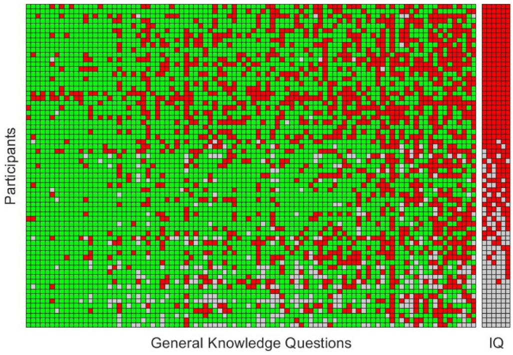
Results

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 1. The figure shows the heterogeneous pattern of participant responses and the distribution of question difficulties. Participants had an overall accuracy of 73.1% on the 94 general knowledge questions when they opted in. Participants opted in to those questions 93.5% of the time.

Figure 1
Participant Responses to Each Question



Note. Green and red colors indicate a correct and incorrect responses, respectively. Gray colors indicate that the participant opted-out of that question. 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 they answered. See the online article for the color version of this figure.

Impossible Questions

Participants opted out of the impossible questions much more frequently than they opted out of the general knowledge questions (34.8% vs. 6.5%, BF > 100,000 via Bayesian A/B test). We term the number of impossible questions that the participant correctly skipped the Impossible Question Criterion (IQC). Higher IQC indicates a higher level of metacognitive ability. Participants with higher IQC were more accurate on the questions they answered, r = 0.51, BF = 2,505, 95% CI [0.31, 0.66], and more likely to opt out of the 94 general knowledge questions, r = 0.73, BF > 100,000, 95% CI [0.58, 0.82]. Figure 2 shows these relationships.

Selecting Participants to Create More Accurate Crowds

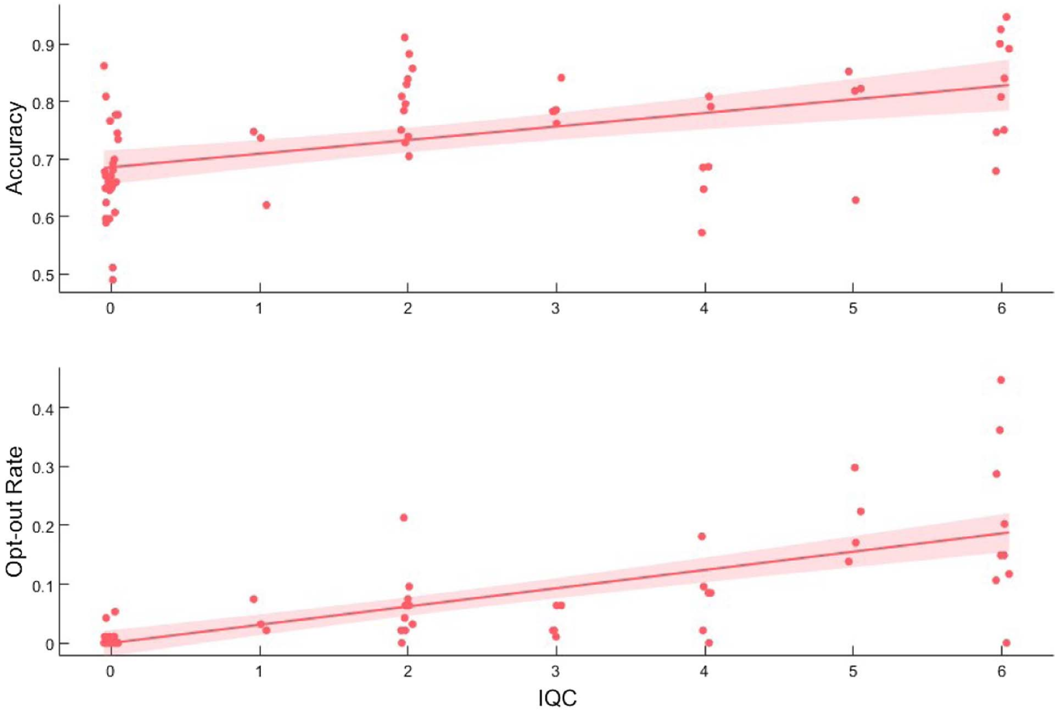
How can impossible questions be leveraged to form more accurate crowds? Since the individuals who correctly opt-out of the impossible questions have higher accuracy, we can use the impossible questions as a filter and only allow those participants who opted-out of a sufficient number of impossible questions into the crowd. Using impossible questions as a filter in this way improves crowd performance relative to an unfiltered crowd when aggregating with average confidence (see Table 2).

We also compute 95% credible intervals to assess how much filtering participants with IQC = 6 improves crowd performance. When scoring with linear loss, the degree of improvement from filtering crowd members in this way is 0.06–0.1 when aggregating with average confidence. When scoring with a Brier score, the 95% credible interval for the improvement ranges from 0.01 to 0.04 when aggregating with average confidence.

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 or seed questions with known answers to select for accurate respondents. 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 average accuracy of 93.5%, with only 22 participants answering any incorrectly. Using these 6 easiest questions as a filter, the resulting 45 person crowd also outperforms the

Figure 2

Opt-Out Rate and Accuracy for the 94 General Knowledge Questions in the Experiment as a Function of the Number of Unanswered Impossible Questions (IQC)



Note. Each point depicts a single participant with some random horizontal displacement for visual clarity. Each line is generated via linear regression and the shaded regions correspond to a 95% confidence interval. See the online article for the color version of this figure.

unfiltered crowd, linear loss: 0.31 versus 0.33, $t(87) = 5.4$, BF = 25,882; Brier score: 0.12 versus 0.13, $t(87) = 3.1$, BF = 8.9.

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 on the 88 questions not used in either filter, linear loss: 0.24 versus 0.31, $t(87) = 6.1$, BF > 100,000; Brier score: 0.10 versus 0.12, $t(87) = 3.1$, BF = 8.8.

The six easiest questions and the six impossible questions are not unusual in their capacity to select for expertise and thereby improve crowd quality. Indeed, most combinations of six questions, when used as a filter, improve the quality of the crowd. We sample 10,000 random combinations of six general knowledge questions and use them to filter out participants as above. We create crowds composed only of those participants who answer all six of the randomly selected questions

Table 2

Crowd Performance Depending on the IQC Used to Select Crowd Members and the Scoring Rule Used to Evaluate Aggregates

IQC	*N*	Linear loss (BF)	Brier score (BF)
0	67	0.31	0.11
2	36	0.27 (>100,000)	0.08 (>100,000)
4	19	0.28 (161)	0.10 (0.32)
6	9	0.18 (>100,000)	0.07 (88)

Note. Bayes factors (BFs) compare the crowd created with IQC to the crowd that includes all participants (denoted with an IQC of 0). All Bayes factors are computed using Bayesian *t*-tests. IQC = Impossible Question Criterion.

correctly. These crowds can be used to determine the extent to which a selection based on impossible questions, in particular, leads to improved crowd performance compared to general knowledge questions (see Figure 3). Filters based on IQC outperform filters based on general knowledge questions 80% of the time when evaluated with linear loss and 73% of the time when evaluated with a Brier score. For a more fine-grained analysis that uses a mixture of general knowledge and impossible questions to filter crowds, see Appendix C.

Since there is a positive correlation between IQC, accuracy, and the number of questions that participants skip, there is a possibility that the benefits of selecting crowd members on the basis of IQC are simply because those participants skipped more questions in general. To assess this, we compare the crowd created by requiring that participants skip all impossible questions (i.e., the crowd with IQC = 6) with the crowd created by requiring that participants skip some number of general knowledge questions. How many questions do we require that participants skip? To eliminate the possibility of bias, we evaluate all possible values for this requirement. We find that regardless of the number of questions we require participants to skip in order to be included in the crowd, the crowd created by selecting participants who skip impossible questions has a lower linear loss and Brier score. The *smallest* benefit of using IQC occurs when we require that participants skip 18 questions when scoring with linear loss, 0.24 versus 0.18, $t(93) = 5.0$, BF = 7,410, and skip three questions when scoring with a Brier score, 0.13 versus 0.08, $t(93) = 2.5$, BF = 2.2.

Discussion

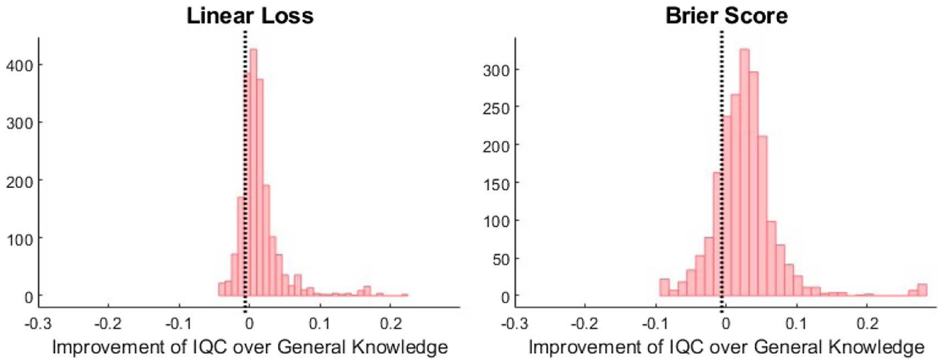
We find that crowd members who skip impossible questions exhibit greater accuracy than their peers. Moreover, the crowd composed only of those participants who skipped impossible questions outperforms the crowd with all individuals. This benefit is greater than using a comparable method to identify experts with general knowledge questions; the metacognitive judgment to skip impossible questions may be a better indicator of expertise than domain knowledge exhibited by answering questions for which there is a correct answer. This highlights the potential for using impossible questions in crowd-sourcing contexts in which crowd members opt-in.

Experiment 2: Reanalysis of Bereby-Meyer et al., 2003

Bereby-Meyer et al. (2003) evaluated how scoring rules framed in terms of gains or losses impact test-taking strategies. The test allowed participants to skip (i.e., opt-out of) any number of multiple-choice questions on a test that

Figure 3

Histograms Showing the Degree of Improvement From Using Impossible Questions Instead of General Knowledge Questions to Identify High-Performing Crowd Members



Note. Positive values indicate that crowds composed of participants who correctly withheld answers from all six impossible questions (IQC) outperformed crowds composed of participants who correctly answered six randomly selected questions (General Knowledge). Each plot corresponds to a different method of scoring crowds. The dotted lines at the 0 point correspond to no difference between resulting crowds. See the online article for the color version of this figure.

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.

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 six 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 the previous experiment.

These data provide the opportunity to immediately replicate our findings from Experiment 1 with a new population and a new set of questions to verify that the main finding replicates across domains. Does the crowd composed of individuals who omitted the unsolvable items once again outperform the crowd that includes all participants?

Aggregating and Scoring Crowds

To the extent that it is possible, we will use the same method of aggregating and scoring crowds as in Experiment 1. However, Bereby-Meyer et al. only collected the accuracy of participants' responses. Since they did not collect confidence ratings, aggregating via average confidence is not possible. Instead, we will aggregate responses by computing the proportion of participants who endorse the correct response option and assigning that probability to the aggregate. We term this equal weighting, which would be equivalent to aggregating with average confidence if all participants had given 100% confidence in their answer. As in Experiment 1, aggregates will be scored with both linear loss and Brier score.

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 of the impossible questions while 8 participants opted out of all 6. The number of impossible questions that a participant skipped (IQC) was positively correlated with accuracy, $r = 0.33$, BF = 22, 95% CI [0.13, 0.50], and the rate at which they opted-out of the general knowledge questions, $r = 0.86$, BF > 100,000, 95% CI [0.78, 0.90].

As in Experiment 1, 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 3). Filtering out respondents that exhibit low metacognitive ability by enforcing a minimum IQC generally results in better crowd performance.

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 populations and questions.

Table 3

Crowd Performance for Bereby-Meyer et al. (2003) Depending on the IQC Used to Select Crowd Members and the Aggregation Method Used to Combine Their Responses

IQC	Aggregation method	*N*	Linear loss (BF)	Brier score (BF)
0	Equal weighting	92	0.43	0.21
2	Equal weighting	35	0.38 (283)	0.17 (16)
4	Equal weighting	28	0.38 (50)	0.17 (4.5)
6	Equal weighting	8	0.35 (1.3)	0.19 (0.23)

Note. Bayes factors compare the crowd created with IQC to the crowd that includes all participants (denoted with an IQC of 0). All Bayes factors are computed using Bayesian *t*-tests. IQC = Impossible Question Criterion.

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 and a reanalysis of data from Bereby-Meyer et al. (2003). Moreover, filters based on impossible questions outperformed most other sets of general knowledge control questions in identifying experts, demonstrating that metacognitive ability can be a better predictor of expertise than historical accuracy.

How should we interpret a participant's decision to skip impossible questions? We have posited that this is a metacognitive measure, but other psychological factors likely influence this decision. For instance, participants who are inclined to take more risks go on to answer questions for which they have a lower degree of confidence (Alnabhan, 2002; Campbell et al., 2004). Additionally, participants may skip questions to reduce the total effort expended on the task (Bogner & Landrock, 2016; Krosnick et al., 1996). Note that in the present study participants who skipped questions exhibited higher accuracy and made greater contributions to the crowd, so effort reduction may not be the right framework for understanding skipping behavior in this experimental setup. Nonetheless, future research could identify the relationship between a participant's propensity to take risks, satisfice by minimizing the effort required to complete the task, and skip questions they do not believe they can answer correctly.

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 when crowd members opt-in (for an analysis comparing Impossible Questions to meta-d' in the current experiment, see Appendix B). A more explicit treatment of the interrelationship between these metacognitive measures would require a mixture of forced-choice questions (which are typically used when estimating meta-d' or measures of overconfidence) and questions that the participants can elect to skip (which are required to assess a participant's ability to skip impossible questions). This would make it possible to clearly differentiate between the cognitive and metacognitive judgments made by participants.

It is unclear how incentives might interact with participants' metacognitive decisions to select questions in an experimental context like the present one. The Bereby et al.'s data provide an example of how researchers can encourage accurate responses by evaluating the relative performance of participants against one another (participants received extra credit if they scored in the top 50% of respondents). However, a more typical method of incentivizing participants does not compare the relative performance of participants. Instead, accuracy is incentivized directly, with higher accuracy resulting in greater payouts regardless of the performance of other participants. This might be done by selecting some questions at random that provide a bonus if answered correctly (e.g., Crump et al., 2013). Such methods are difficult to implement when participants are permitted to opt-out of questions. If participants receive no bonus for skipping questions, then skipping will be discouraged altogether. If instead only answered questions are eligible for bonuses, participants will be incentivized to employ the strictest possible criteria for selecting questions and therefore skip many more questions than they might otherwise. When trying to aggregate many responses to create an accurate crowd as we did in this study, both of these behaviors would likely have deleterious effects on the resulting crowds by either reducing the size of the crowd or discouraging crowd members from skipping questions when they lack relevant knowledge. Future research could address how to leverage incentives in an opt-in setting to improve crowd performance by aligning participants' incentives with the goals of the crowd.

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 (Budescu & Chen, 2015) 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. Existing methods of estimating expertise in contexts where the ground truth is unknown rely on the similarity of responses between individuals (e.g., Kurvers et al., 2019). However, these methods require that respondents answer all questions. Our method is similar to that of Palley and Soll (2019) and Palley and Satopää (2021) in that metacognitive measures allow us to arrive at estimates of expertise even when there are few questions or participants answer few questions.

Much of the existing research 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 underexplored area of research because it relates to crowd wisdom in real-world applications. Indeed, many existing methods that improve the accuracy of crowds already implicitly rely on metacognition. Confidence-weighted pooling exploits the relationship between the metacognitive judgment of confidence and item-level accuracy (Qyama et al., 2013), the Surprisingly Popular Algorithm exploits the relationship between meta-knowledge of others' beliefs and accuracy (Prelec et al., 2013, 2017), and the benefit of opting-in is due to participants' ability to recognize their own knowledge (Bennett et al., 2018). Making this connection explicit highlights the need for further research on role of metacognition in crowd contexts.

References

Alnabhan, M. (2002) An empirical investigation of the effects of three methods of handling guessing and risk taking on the psychometric indices of a test. *Social Behavior and Personality: An International Journal*, *30*(7), 645–652. https://doi.org/10.2224/sbp.2002.30.7.645

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., Wallsten, T. S., & Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*(2), 130–147. https://doi.org/10.1037/1076-898X.6.2.130

Atanasov, P., Rescober, P., Stone, E., Swift, S. A., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2017). Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, *63*(3), 691–706. https://doi.org/10.1287/mnsc.2015.2374

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When knowledge knows no bounds: Self-perceived expertise predicts claims of impossible knowledge. *Psychological Science*, *26*(8), 1295–1303. https://doi.org/10.1177/0956797615588195

Beger, A., & Ward, M. D. (2019). *Assessing Amazon Turker and automated machine forecasts in the hybrid forecasting competition* [Conference session]. 7th annual asian political methodology conference, Kyoto, Japan.

Bennett, S. T., Benjamin, A. S., Mistry, P. K., & Steyvers, M. (2018). Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. *Computational Brain anf Behavior*, *1*(1), 90–99. https://doi.org/10.1007/s42113-018-0006-4

Bensch, D., Paulhus, D. L., Stankov, L., & Ziegler, M. (2019). Teasing apart overclaiming, overconfidence, and socially desirable responding. *Assessment*, *26*(3), 351–363. https://doi.org/10.1177/1073191117700268

Bereby-Meyer, Y., Meyer, J., & Budescu, D. V. (2003). Decision making under internal uncertainty: The case of multiple-choice tests with different scoring rules. *Acta Psychological*, *12*(2), 207–220. https://doi.org/10.1016/S0001-6918(02)00085-9

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J., & Bennett, S. E. (1980). Pseudo-Opinions on public affairs. *Public Opinion Quarterly*, *44*(2), 198–209. https://doi.org/10.1086/268584

Bishop, G. F., Tuchfarber, A. J., & Oldendick, R. W. (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly*, *50*(2), 240–250. https://doi.org/10.1086/268978

Bogner, K., & Landrock, U. (2016). *Response biases in standardised surveys*. GESIS survey guidelines.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280. https://doi.org/10.1287/mnsc.2014.1909

Burnap, A., Ren, Y., Gerth, R., Papazoglou, G., Gonzalez, R., & Papalambros, P. Y. (2015). When crowdsourcing fails: A study of expertise on crowdsourced design evaluation. *Journal of Mechanical Design*, *137*(3), Article 031101. https://doi.org/10.1115/1.4029065

Campbell, W. K., Goodie, A. S., & Foster, J. D. (2004). Narcissism, confidence, and risk attitude. *Journal of Behavioral Decision Making*, *17*(4), 297–311. https://doi.org/10.1002/bdm.475

Castano, S., Ferrara, A., & Montanelli, S. (2019). Leveraging crowd skills and consensus for collaborative web-resource labeling. *Future Generation Computer Systems*, *95*, 790–801. https://doi.org/10.1016/j.future.2017.12.024

Chen, E., Budescu, D. V., Lakshmikanth, S. K., Mellers, B. A., & Tetlock, P. E. (2016). Validating the contribution-weighted model: Robustness and cost-benefit analyses. *Decision Analysis*, *13*(2), 128–152. https://doi.org/10.1287/deca.2016.0329

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's mechanical Turk as a tool for experimental behavioral research. *PLOS ONE*, *8*(3), Article e57410. https://doi.org/10.1177/0963721414531598

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79–101. https://doi.org/10.1037/dec0000004

Dunlop, P. D., Bourdage, J. S., de Vries, R. E., McNeill, I. M., Jorritsma, K., Orchard, M., Austen, T., Baines, T., & Choe, W.-K. (2020). Liar! Liar! (when stakes are higher): Understanding how the overclaiming technique can be used to measure faking in personnel selection. *Journal of Applied Psychology*, *105*(8), 784–799. https://doi.org/10.1037/apl0000463

Eickhoff, C. (2018). *Cognitive biases in crowdsourcing* [Conference session]. Proceedings of the eleventh ACM international conference on web search and data mining, Marina Del Rey, California, United States. https://doi.org/10.1145/3159652.3159654

Fan, Y., Budescu, D. V., Mandel, D., & Himmelstein, M. (2019). Improving accuracy by coherence weighting of direct and ratio probability judgments. *Decision Analysis*, *16*(3), 197–217. https://doi.org/10.1287/deca.2018.0388

Frederick, S. (2005). Cognitive reflection and decision making cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Grushka-Cockayne, Y., Jose, V. R. R., & Lichtendahl, K. C., Jr. (2017). Ensembles of overfit and overconfident forecasts Ensembles of overfit and overconfident forecasts. *Management Science*, *63*(4), 1110–1130. https://doi.org/10.1287/mnsc.2015.2389

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? Are manipulation checks necessary? *Frontiers in Psychology*, *9*, Article 998. https://doi.org/10.3389/fpsyg.2018.00998

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Heim, E., Roß, T., Seitel, A., März, K., Stieltjes, B., Eisenmann, M., Lebert, J., Metzger, J., Sommer, G., Sauter, A. W., Schwartz, F. R., Termer, A., Wagner, F., Kenngott, H. G., & Maier-Hein, L. (2018). Large-scale medical image annotation with crowd-powered algorithms. *Journal of Medical Imaging*, *5*(3), Article 034002. https://doi.org/10.1117/1.JMI.5.3.034002

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

Hill, S., & Ready-Campbell, N. (2011). Expert stock picker: the wisdom of (experts in) crowds. *International Journal of Electronic Commerce*, *15*(3), 73–102. https://doi.org/10.2753/JEC1086-4415150304

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), Article 2. https://doi.org/10.7771/1932-6246.1167

JASP Team. (2020). *JASP (Version 0.12)* [Computer software]. https://jasp-stats.org/

Johanna, B. I., & van der Heijden, M. (2000). The development and psychometric evaluation of a multidimensional measurement instrument of professional expertise. *High Ability Studies*, *11*(1), 9–39. https://doi.org/10.1016/j.ics.2005.02.061

Juni, M. Z., & Eckstein, M. P. (2017). The wisdom of crowds for visual search. *Proceedings of the National Academy of Sciences*, *114*(21), E4306–E4315. https://doi.org/10.1073/pnas.1610732114

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Klusmann, V., Evers, A., Schwarzer, R., & Heuser, I. (2011). A brief questionnaire on metacognition: Psychometric properties. *Aging and Mental Health*, *15*(8), 1052–1062. https://doi.org/10.1080/13607863.2011.583624

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, *1996*(70), 29–44. https://doi.org/10.1002/ev.1033

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Moussaid, M., & Argenziano, G. (2019). Wolf, M. How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, *5*(11), Article eaaw9011. https://doi.org/10.1126/sciadv.aaw9011

Li, Q., & Varshney, P. K. (2017). *Does confidence reporting from the crowd benefit crowdsourcing performance?* [Conference session]. Proceedings of the 2nd international workshop on social sensing, Pittsburgh, United States. https://doi.org/10.1145/3055601.3055607

Mandel, D. R. (2008). Violations of coherence in subjective probability: A representational and assessment processes account. *Cognition*, *106*(1), 130–156. https://doi.org/10.1016/j.cognition.2007.01.001

Maniscalco, B. (2020). *Type 2 signal detection theory analysis using meta-d*. Retrieved July 27, 2020, from http://www.columbia.edu/~bsm2105/type2sdt/

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Maniscalco, B., & Lau, H. (2014). 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* (pp. 25–66). Springer. https://doi.org/10.1007/978-3-642-45190-4_3

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276–299. https://doi.org/10.1037/a0036677

Martire, K. A., Growns, B., & Navarro, D. J. (2018). What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts. *Psychonomic Bulletin and Review*, *25*(6), 2346–2355. https://doi.org/10.3758/s13423-018-1448-3

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, *21*(1), 1–14. https://doi.org/10.1037/xap0000040

Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2017). A neglected dimension of good forecasting judgment: The questions we choose also matter. *International Journal of Forecasting*, *33*(4), 817–832. https://doi.org/10.1016/j.ijforecast.2017.04.002

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

Palley, A., & Satopää, V. (2021). *Boosting the wisdom of crowds within a single judgment problem: Weighted averaging based on peer predictions*. Social Science Research Network. https://doi.org/10.2139/ssrn.3504286

Palley, A. B., & Soll, J. B. (2019). Extracting the wisdom of crowds when information is shared. *Management Science*, *65*(5), 2291–2309. https://doi.org/10.1287/mnsc.2018.3047

Peeters, T. (2018). Testing the wisdom of crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting*, *34*(1), 17–29. https://doi.org/10.1016/j.ijforecast.2017.08.002

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking *Journal of personality*, *88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Pion-Tonachini, L., Makeig, S., & Kreutz-Delgado, K. (2017). Crowd labeling latent Dirichlet allocation. *Knowledge and Information Systems*, *53*(3), 749–765. https://doi.org/10.1007/s10115-017-1053-1

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

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535. https://doi.org/10.1038/nature21054

Quigley, J., Colson, A., Aspinall, W., & Cooke, R. M. (2018). Elicitation in the classical model Elicitation in the classical model. In *Elicitation* (pp. 15–36). Springer. https://doi.org/10.1007/978-3-319-65052-4_2

Rouder, J. N. (2014). *Bayes factor calculators*. Retrieved July 27, 2020, from http://pcl.missouri.edu/bayesfactor.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis *Psychonomic Bulletin and Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Stastny, B. J., & Lehner, P. E. (2018). Comparative evaluation of the forecast accuracy of analysis reports and a prediction market. *Judgment and Decision Making*, *13*(2), 202–211.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V., & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, *95*(3), 261–289. https://doi.org/10.1007/s10994-013-5401-4

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin and Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Welinder, P., Branson, S., Perona, P., & Belongie, S. J. (2010). The multidimensional wisdom of crowds. In *Advances in neural information processing systems* (pp. 2424–2432).

Witkowski, J., Atanasov, P., Ungar, L., & Krause, A. (2017). Proper proxy scoring rules. *Proceedings of the AAAI conference on artificial intelligence* (Vol. *31*).

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of economic perspectives*, *18*(2), 107–126. https://doi.org/10.1257/0895330041371321

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science*, *36*(3), 452–470. https://doi.org/10.1111/j.1551-6709.2011.01223.x

(*Appendices follow*)

Appendix A

Aggregating With Majority Rule

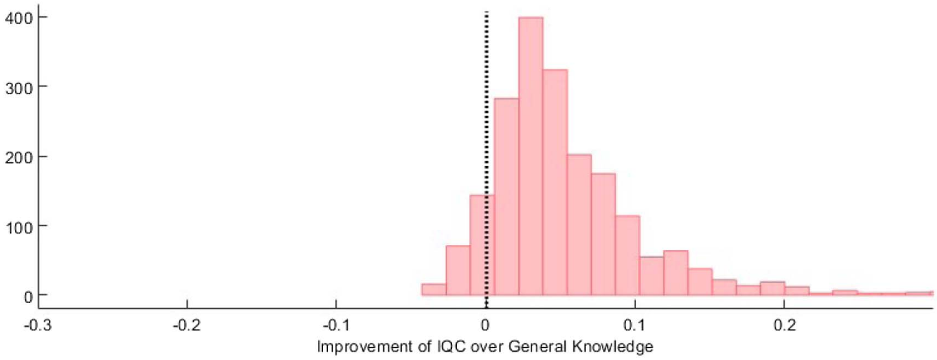
In this section, we report results of aggregating with majority rule rather than average confidence. The majority rule aggregate is equal to the most common response among crowd members. Since aggregating this way does not produce a probability estimate of each response option, linear loss and Brier score are not valid methods of evaluating the aggregate. Instead, the aggregate is scored on the basis of the proportion of questions for which it provides the correct answer. As an example of aggregating with majority rule, consider a crowd of four participants wherein one participant opts out, one participant answers incorrectly, and two participants answer correctly. In this case, the majority rule would produce the correct answer since most of the participants who answered the question were correct.

We report the results from the major analyses conducted in Experiment 1 using majority rule to aggregate crowds instead of average confidence. The crowd composed of all individuals in Experiment 1 and aggregated with majority rule produces the correct answer 86% of the time. When aggregating only the responses from those participants who skipped all six impossible questions, the majority rule aggregate produces the correct answer 91% of the time. On the basis of a Bayesian A/B test, it is ambiguous if this crowd outperforms the one which includes all participants (86% vs. 91%, BF = 0.73).

We also reproduced the analysis wherein we sampled general knowledge questions to use as seed questions. As before, we sampled six questions at random 10,000 times to use as seed questions, and compared the crowd of participants who answered those six questions correctly with the crowd that is filtered based on IQC. In this analysis, the filters based on IQC outperform those based on general knowledge questions 92% of the time (see Figure A1).

Figure A1

Histogram Showing the Degree of Improvement From Using Impossible Questions Instead of General Knowledge Questions to Identify High-Performing Crowd Members



Note. Crowds are aggregating with majority rule. Positive values indicate that crowds composed of participants who correctly withheld answers from all six impossible questions (IQC) outperformed crowds composed of participants who correctly answered six randomly selected questions (General Knowledge). The dotted line corresponds to no difference between resulting crowds. See the online article for the color version of this figure.

Appendix B

Confidence as a Metacognitive Measure in an Opt-In Context

Confidence ratings can be used to compute measures of metacognitive ability such as *meta-d'*. 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.

(*Appendices continue*)

Indeed, in a crowd-sourcing context in which participants could opt-in, there was no benefit to crowd performance when leveraging confidence ratings (Li & Varshney, 2017). 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 this other metacognitive measure.

We computed meta-d' and the ratio between meta-d' and d' with publicly available software (Maniscalco, 2020; Maniscalco & Lau, 2012, 2014). To compute meta-d', we converted the scalar ratings solicited in our experiment into categorical ratings. We did this by treating all confidence ratings above the grand median confidence rating (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) may be spurious. 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.

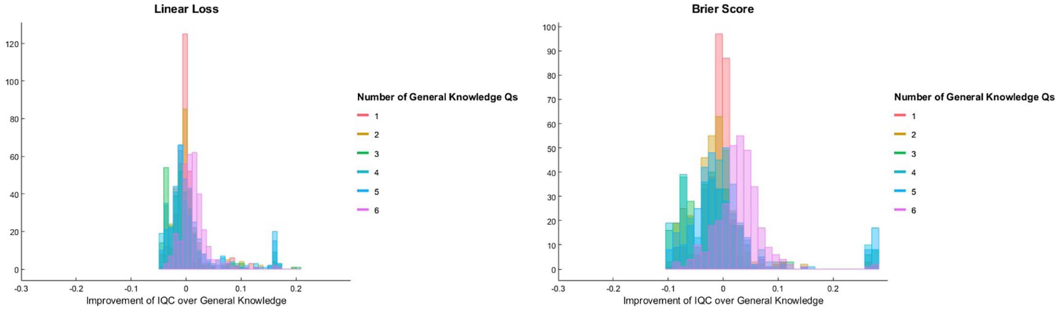
Appendix C

Mixing General Knowledge and Impossible Questions as Seed Questions

We investigated whether varying the proportion of seed questions that came from impossible versus possible questions impacted resulting crowd performance. To do this, we created crowds by randomly selecting a combination of general knowledge and impossible questions. Participants were excluded if they answered any of those general knowledge questions incorrectly or answered one of the selected impossible questions. Crowd performance of these mixed crowds is compared to the crowd created only using impossible questions (see Figure C1). There may be some benefit from using a mixture of general knowledge questions and impossible questions to select crowd members; the average performance of crowds created using three general knowledge questions and three impossible questions has the best average performance across scoring rules.

Figure C1

Histograms Showing the Relative Performance of Crowds Depending on the Method Used to Identify Experts



Note. Each color compares a filter composed of a mixture of general knowledge and impossible questions with the filter composed exclusively of impossible questions (i.e., the crowd corresponding to an IQC of six). Positive values indicate that the crowd created with impossible questions outperformed the crowd created with a mixture of general knowledge and impossible questions (and so negative values indicate that the mixed crowd outperformed the crowd based on IQC alone). Each plot corresponds to a different method of scoring crowds. See the online article for the color version of this figure.

Received September 21, 2020
Revision received November 3, 2021
Accepted November 4, 2021 ■