Deliberation increases the wisdom of crowds

Joaquin Navajas¹, Tamara Niella², Gerry Garbulsky³, Bahador Bahrami¹ & Mariano Sigman²

¹ Crowd Cognition Group, Institute of Cognitive Neuroscience, University College London, United Kingdom
² Universidad Torcuato Di Tella, Buenos Aires, Argentina
³ TED, Buenos Aires, Argentina

The aggregation of many independent estimates can outperform the most accurate individual judgment. This centenarian finding, popularly known as the “wisdom of crowds”, has recently been applied to problems ranging from the diagnosis of cancer to financial forecasting. It is widely believed that the key to collective accuracy is to preserve the independence of individuals in a crowd. Contrary to this prevailing view, we show that deliberation and discussion improves collective wisdom. We asked a live crowd (N=5180) to respond to general knowledge questions (e.g. the height of the Eiffel Tower). Participants first answered individually, then deliberated and made consensus decisions in groups of five, and finally provided revised individual estimates. We found that consensus and revised estimates were less biased and more diverse than what a uniform aggregation of independent opinions could achieve. Consequently, the average of different consensus decisions was substantially more accurate than aggregating the independent opinions. Even combining as few as four consensus choices outperformed the wisdom of thousands of individuals. Our results indicate that averaging information from independent debates is a highly effective strategy for harnessing our collective knowledge.
Whether or not humans benefit from collective decision-making has puzzled mankind since the origin of political thought (1). Theoretically, aggregating the opinions of many unbiased and independent agents can outperform the best single judgment (2), which is why crowds are sometimes wiser than their individuals (3, 4). The idea of wise crowds, however, is at odds with the pervasiveness of poor collective judgment (5). Human crowds may fail for two reasons. First, human choices are frequently plagued with numerous systematic biases (6). Second, opinions in a crowd are rarely independent. Social interactions often cause informational cascades, which correlate opinions, aligning and exaggerating the individual biases (7). This imitative behavior may lead to herding (5), financial bubbles (8), rich-get-richer dynamics (9, 10), zealotry (11), and violence contagion (12). Empirical research has shown that even weak social influence can undermine the wisdom of the crowd (13), and that collectives are less biased when their individuals resist peer influence (14). Extensive evidence suggests that the key to collective intelligence is to protect the independence of opinions within a group.

However, in many of those previous works, social interaction was operationalized by participants observing others’ choices without discussing them. These reductionist implementations of social influence may have oversimplified collective behavior. For example, human groups can communicate their uncertainty and make joint decisions that reflect the reliability of each group member (15). During peer discussion, people also exchange shareable arguments (16, 17), which promote the understanding of a problem (18). Groups can reach consensuses that are outside the span of their individual decisions (18, 19), even if a minority (20) or no one (18) knew the correct answer before interaction. These findings lead to the following questions: can crowds be any wiser if they debated their choices? Should their members be kept as independent as possible and aggregate their uninfluenced, individual opinions? 110 years after Galton’s seminal study (3), we addressed these questions by performing an experiment on a large live crowd (Fig 1A).
We asked the members of audience (N=5180, 2468 female, aged 30.1±11.6 years) attending a popular TEDx event in Buenos Aires (Argentina) to answer eight questions involving approximate estimates to general knowledge quantities (e.g., what is the height in meters of the Eiffel Tower? c.f. Methods). Each participant was provided with pen and an answer sheet linked to their seat number. The event’s speaker (author M.S.) conducted the crowd from the stage (Fig. 1A). In the first stage of the experiment, the speaker asked eight questions (Supplementary Table 1) and gave participants 20 seconds to respond to each of them (stage i1, left panel in Fig. 1A). Then, participants were instructed to organize into groups of five based on a numerical code in their answer sheet (see Methods). The speaker repeated four of the eight questions and gave each group one minute to reach a consensus (stage c, middle panel in Fig. 1A). Finally, the eight questions were presented again from stage and participants had 20 seconds to write down their individual estimate, which gave them a chance to revise their opinions (stage i2, right panel in Fig. 1A). In all individual responses participants also reported their confidence in a scale from 0 to 10.

Responses to different questions were distributed differently. To pull the data across questions, we used a non-parametric normalizing method, used for rejecting outliers (21) (see Methods). Normalizing allowed us to visualize the grouped data parsimoniously, but all our main findings are independent of this step (Supplementary Fig. 1). As expected, averaging the initial estimates from n participants led to a significant decrease in collective error as n increased (F(4,999)= 477.3, p<0; blue lines in Fig. 1B), replicating the classic wisdom-of-crowd effects (3). The average of all initial opinions in the auditorium (N=5180) led to 52% error reduction compared to the individual estimates (Wilcoxon sign rank test, z=61.79, p<0).

We then focused on the effect of debate on the wisdom of crowds, and studied whether social interaction and peer discussion impaired (13, 14) or promoted (17, 18) collective wisdom. To disentangle these two main alternative hypotheses, we looked at the consensus estimates. We randomly sampled m groups and compared the wisdom of m consensus estimates (stage c) against the wisdom of n initial opinions (stage i1, n = 5m since there were
5 participants on each group). This analysis is based on the 280 groups (1400 participants) that had valid data from all of their members (see Methods). We observed that the average of as few as 3 collective estimates was more accurate than the mean of the 15 independent initial estimates (blue line at \( n = 15 \) vs. black line at \( m = 3 \) in Fig. 1B, \( z=13.25, p=10^{-40} \)). The effect was even more clear when comparing 4 collective choices against the 20 individual decisions comprising the same 4 groups (blue line at \( n = 20 \) vs. black line at \( m = 4 \) in Fig. 1B, \( z=20.79, p=10^{-98} \)). Most notably, the average of 4 collective estimates was even more accurate (by 49.2% reduction in error) than the average of the 1400 initial individual estimates (blue data point at \( n = 1400 \) vs. black line at \( m = 4 \) in Fig 1B, \( z=13.92, p=10^{-44} \)). In principle, this could simply result from participants having a second chance to think about these questions, and providing more accurate individual estimates to the group discussion than the ones initially reported. However, our data rules out this possibility since one or two collective estimates are as accurate as 5 or 10 independent initial estimates respectively (Fig. 1B). In other word this is the result of a crowd of crowds.

Participants used the chance to change their minds after interaction and this reduced their individual error (mean error reduction of 31%, \( z=19.16, p=10^{-82} \)). More importantly, revised estimates gave rise to greater wisdom of crowds compared to initial estimates (blue line vs. red line in Fig. 1B, \( F(1,999)= 4458.6, p<0 \)). When compared to collective choices, the average of \( n \) revised decisions was overall more accurate than the average of \( m \) group decisions (black line vs. red line in Fig. 1B, \( F(1,999)= 2510.4, p<0 \), although this depended on the specific question asked (interaction \( F(3,999)= 834.7, p<0 \); see Supplementary Fig. 1). Taken together, these findings are the first demonstration that face-to-face social interaction brings remarkable benefits in accuracy and efficiency to the wisdom of the crowds. These results raise the question of how social interaction, which is expected to instigate herding (7), could have improved collective estimates.

Several observations about the bias and variance of the distributions of estimates help understanding our results. We found that the consensus decisions were less biased than the
average of initial estimates ($z=2.15$, $p=0.03$, Supplementary Fig. 2). When participants changed their mind, they approached the (less biased) consensus: revised opinions became closer to the consensus than to the average of initial answers ($z=27.15$, $p=10^{-162}$). This indicates that deliberation led to a better consensus than what a simple averaging procedure (with uniform weights) could achieve. Moreover, in line with previous reports that social influence reduces the diversity of opinions (13, 14) we found that, within each group, revised responses converged towards each other: the variance of revised estimates within each group was smaller than the variance of the initial estimates (Wilcoxon sign rank test of the variance of responses on each group before vs after interaction, $z=18.33$, $p=10^{-75}$). However, interaction actually increased the variance of responses between groups: the distribution of the average of initial estimates (obtained by averaging stage $i1$ estimates on each group) had less variance than the average of revised estimates (obtained by averaging stage $i2$ estimates on each group, squared rank test for homogeneity, $p<0.01$). Previous research in social psychology also found a similar effect; consensus decisions are typically more extreme than the average individual choice, a phenomenon known as ‘group polarization’ (22).

Previous studies have proposed that a fundamental condition to elicit the wisdom-of-crowds effect is the diversity of opinions (4, 23). Because we saw that interaction decreased the variance of estimates within groups but increased the variance between groups, we reasoned that sampling opinions from different groups might bring even larger benefits to the crowd. To test this idea, we sampled our population in two ways to test the impact of within- and between-group variance on the wisdom of crowds (Fig. 1C). In the within-groups condition, we sampled $n$ individuals coming from $m = n/5$ different groups. This was the same sampling procedure that we used in Fig. 1B. In the between-groups sampling, we selected $n$ individuals, each coming from a different group. Because different groups were randomly placed in different locations in the auditorium, we expected that sampling between-groups would break the effect of local correlations, and decrease the collective error.
Fig.1. Wisdom of interacting crowds. (A) A live crowd (N=5180) answered general knowledge questions in three stages. Left: initial individual estimate (stage $i_1$). Middle: consensus (stage c). Right: revised individual estimate (stage $i_2$). In stage c, a moderator (white) recorded the group’s consensus estimate. (B) Normalized error of the average of $n$ individual answers (blue line for stage $i_1$, red line for stage $i_2$), and normalized error of the average of $m = n/5$ collective estimates (black line, stage c). Bars are s.e.m. (C) We aggregated the individual estimates in two different ways: either by sampling participants all from the same interacting groups (within-groups condition) or by sampling each participant from a different interacting group (between-groups condition). The insets sketch these two conditions; participants shaded by the same color were averaged together. Normalized error of the average of $n$ individual answers at stage $i_2$ for the within-groups condition (solid line) and the between-groups condition (dashed line).
Consistent with our predictions, we found that breaking the local correlations by between-group sampling led to a large error reduction (red solid line vs. red dashed line in Fig. 1C, 26% error reduction on average, $F(1,999)=25824.1, p ~ 0$). In fact, averaging only 5 revised estimates coming from 5 different groups outperformed the aggregation of all initial independent decisions in the auditorium ($z=25.91, p = 10^{-148}$). Adding more decisions using this sampling procedure led to a significant decrease in error ($F(4,999)=249.34, p ~ 0$). Aggregating revised estimates from different randomly sampled groups was a highly effective strategy to improve collective accuracy and efficiency, even with a very small number of samples.

Our results are in contrast with an extensive literature on herding (7) and dysfunctional group behavior (24), which exhorts us to remain as independent as possible. Instead, our findings are consistent with research in collaborative learning showing that “think-pair-share” strategies (25) and peer discussion (18) can increase the understanding of conceptual problems. However, these findings offer a key novel insight largely overlooked in the literature on aggregation of opinions: pooling together collective estimates made by independent, small groups that interacted within themselves increases the wisdom-of-crowds effect. The potential applications of this approach are numerous and range from improving structured communication methods that explicitly avoid face-to-face interactions (26), to the aggregation of political and economic forecasts (27) and the design of wiser public policies (28).

The first study on the wisdom of crowds was regarded as an empirical demonstration that democratic aggregation rules can be trustworthy and efficient (3). Since then, all attempts to increase collective wisdom have been based on the idea that some decisions have larger merit than others and pursued some ideal non-uniform weighting algorithm (14, 29, 30). For example, previous studies proposed to average the responses of a ‘select crowds’ defined by higher confidence (14) or expertise (29), or to select the ‘surprisingly popular’ minority answers (30). Although these methods lead to substantial improvements in performance, implementing simple majority rules may still be preferred for other reasons which may include sharing responsibility (31), promoting social inclusion (32), and avoiding elitism or inequality (33).
Here, we showed that the wisdom of crowds can be increased by simple face-to-face discussion within groups coupled with between-group sampling. Critically, this is achieved without compromising the democratic principle of ‘one vote, one value’ (34). The idea that deliberation gives legitimacy and potentially improves democratic choices is not a new one (35). But, to our knowledge, this is the first empirical demonstration that a deliberative democratic procedure increases the wisdom of crowds. Our results support political theories postulating that authentic deliberation, and not simply voting, can lead to better democratic choices (36).

References

1. B. Jowett, The Republic of Plato (Clarendon press, Oxford, 1888).

2. M. Condorcet, Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix (L'impremerie royale, Paris, France, 1785).

3. F. Galton, Vox populi. Nature. 7, 450-451 (1907).

4. J. Surowiecki, The Wisdom of Crowds (Little, Brown, London, UK, 2004).

5. C. MacKay, Extraordinary Popular Delusions the Madness of Crowds (Wordsworth Editions Limited, Ware, UK, ed. 4th, 1841).

6. A. Tversky, D. Kahneman, Judgment under Uncertainty: Heuristics and Biases. Science. 185, 1124-1131 (1974).

7. R. M. Raafat, N. Chater, C. Frith, Herding in humans. Trends Cogn. Sci. 13, 420-428 (2009).

8. V. V. Chari, P. J. Kehoe, Financial crises as herds: overturning the critiques. J. Econ. Theory. 119, 128-150 (2004).

9. M. J. Salganik, P. S. Dodds, D. J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market. Science. 311, 854-856 (2006).

10. L. Muchnik, S. Aral, S. J. Taylor, Social influence bias: a randomized experiment. Science. 341, 647-651 (2013).

11. L. Festinger, H. W. Riecken, S. Schachter, When prophecy fails: A social and psychological study of a modern group that predicted the end of the world. (1956).

12. L. Berkowitz, J. Macaulay, The contagion of criminal violence. Sociometry., 238-260 (1971).
13. J. Lorenz, H. Rauhut, F. Schweitzer, D. Helbing, How social influence can undermine the wisdom of crowd effect. Proc. Natl. Acad. Sci. U. S. A. **108**, 9020-9025 (2011).

14. G. Madirolas, G. G. de Polavieja, Improving Collective Estimations Using Resistance to Social Influence. PLoS Comput. Biol. **11**, e1004594 (2015).

15. B. Bahrami et al., Optimally interacting minds. Science. **329**, 1081-1085 (2010).

16. H. Mercier, D. Sperber, Why do humans reason? Arguments for an argumentative theory. Behav. Brain Sci. **34**, 57-74 (2011).

17. H. Mercier, D. Sperber, "Two heads are better" stands to reason. Science. **336**, 979-a (2012).

18. M. K. Smith et al., Why peer discussion improves student performance on in-class concept questions. Science. **323**, 122-124 (2009).

19. P. R. Laughlin, B. L. Bonner, A. G. Miner, Groups perform better than the best individuals on letters-to-numbers problems. Organ. Behav. Hum. Decis. Process. **88**, 605-620 (2002).

20. D. M. M. Geil, Collaborative reasoning: Evidence for collective rationality. Thinking & Reasoning. **4**, 231-248 (1998).

21. C. Leys, C. Ley, O. Klein, P. Bernard, L. Licata, Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. J. Exp. Soc. Psychol. **49**, 764-766 (2013).

22. D. G. Myers, H. Lamm, The group polarization phenomenon. Psychol. Bull. **83**, 602 (1976).

23. L. Hong, S. E. Page, Groups of diverse problem solvers can outperform groups of high-ability problem solvers. Proc. Natl. Acad. Sci. U. S. A. **101**, 16385-16389 (2004).

24. S. E. Asch, Opinions and Social Pressure. Sci. Am. **193**, 31-35 (1955).

25. F. T. Lyman, The responsive classroom discussion: The inclusion of all students (University of Maryland Press, Maryland, 1981), pp. 113.

26. N. Dalkey, O. Helmer, An experimental application of the Delphi method to the use of experts. Management Science. **9**, 458-467 (1963).

27. P. Tetlock, Expert political judgment: How good is it? How can we know? (Princeton University Press, , 2005).

28. C. R. Sunstein, Infotopia: How many minds produce knowledge (Oxford University Press, , 2006).

29. A. E. Mannes, J. B. Soll, R. P. Larrick, The wisdom of select crowds. J. Pers. Soc. Psychol. **107**, 276 (2014).

30. D. Prelec, H. S. Seung, J. McCoy, A solution to the single-question crowd wisdom problem. Nature. **541**, 532-535 (2017).
31. N. Harvey, I. Fischer, Taking advice: Accepting help, improving judgment, and sharing responsibility. Organ. Behav. Hum. Decis. Process. **70**, 117-133 (1997).

32. N. I. Eisenberger, M. D. Lieberman, K. D. Williams, Does rejection hurt? An FMRI study of social exclusion. Science. **302**, 290-292 (2003).

33. A. Mahmoodi et al., Equality bias impairs collective decision-making across cultures. Proc. Natl. Acad. Sci. U. S. A. **112**, 3835-3840 (2015).

34. F. Galton, One vote, one value. Nature. **75**, 414 (1907).

35. J. S. Mill, On liberty (John W Parker and Son, London, 1859).

36. J. Bohman, W. Rehg, Deliberative democracy: Essays on reason and politics (MIT press, Boston, MA, 1997).
Methods

Context

The experiment was performed during a TEDx event in Buenos Aires, Argentina (http://www.tedxriodelaplata.org/) on September 24, 2015. This was the third edition of an initiative called TEDxperiments (http://www.tedxriodelaplata.org/tedxperiments), aimed at constructing knowledge on human communication by performing behavioral experiments on large TEDx audiences. The first two editions studied the cost of interruptions on human interaction (37), and the use of a competition bias in a “zero-sum fallacy” game (38).

Materials

Research assistants handled one pen and one A4 paper to each participant. The A4 paper was folded on the long edge and had four pages. On page 1, participants were informed about their group number and their role in the group. The three stages of the experiment (Fig. 1A) could be completed in pages 2, 3, and 4, respectively. On page 4, participants could also complete information about their age and gender.

Experimental procedure

The speaker (author M.S.) announced that his section would consist in a behavioral experiment. Participants were informed that their participation was completely voluntary and they could simply choose not to participate or withdraw their participation at any time. A total of 5180 participants (2468 female, mean age 30.1 years, s.d.: 11.6 years) performed the experiment. All data were completely anonymous. This experimental procedure was approved by the ethics committee of CEMIC (Centro de Educación Médica e Investigaciones Clínicas Norberto Quirno).
Stage i1: individual decisions

The speaker announced that, in the first part of the experiment, participants would make individual decisions. Subjects answered eight general knowledge questions that involved the estimation of an uncertain number (e.g., what is the height in meters of the Eiffel Tower?). Each question (Supplementary Table 1) had one code (e.g. EIFFEL) and two boxes. Participants were instructed to fill the first box with their estimate, and the second box with their confidence in a scale from 0 to 10. Before the beginning of stage i1, the speaker completed one exemplary question in the screen, and then read the eight questions. Participants were given 20 seconds to answer each question.

Stage c: collective decisions

In the second part (stage c), we asked participants to make collective decisions. First, they were instructed to find other members in their group according to a numerical code found in page 1. Each group had six members, and all participants were seated next to each other in two consecutive rows. The speaker announced that there were two possible roles in the group: player or moderator. Each group had five players and one moderator. Each participant could find their assigned role in page 1 (e.g., “You are the moderator in group 765” or “You are a player in group 391”). Players were instructed to reach a consensus and report it to the moderator in a maximum of 60 seconds. Moderators were given verbal and written instructions to not participate nor intercede in the decisions made by the players. The role of the moderators was simply to write down the collective decisions made by the players in their group. Moderators were also instructed to write down an ‘X’ if there was lack of consensus among the group. Groups were asked to answer four of the eight questions from stage i1 (see Supplementary Table 1). The speaker read the four questions again, and announced the moments in which time was over.
Stage i2: Revised decisions

Finally, participants were allowed to revise their individual decisions and confidence. The speaker emphasized that this part was individual, and read all eight questions of stage i1 again.

Data collection and digitalization

At the end of the talk, we collected the papers as participants exited the auditorium. Over the week following the event, five data-entry research assistants digitalized these data using a keyboard. We collected 5180 papers: 4232 players and 946 moderators. Many of these 946 potential groups had incomplete data due to at least one missing player; overall, we collected 280 complete groups. All data reported in Fig. 1 is based on those 280 complete groups (1400 players). For the comparison between individual, collective, and revised estimates, we focus on the four questions answered at stage c.

Non-parametric normalization

The distributions of responses were spread around different values on each question (Supplementary Figure 1). To normalize these distributions, we used a non-parametric approach inspired in the outlier detection literature (21). We calculated the deviance of each data point $x_i$ around the median, and normalized this value by the median absolute deviance

$$n_i = \frac{x_i - \text{median}(x)}{\text{median}(|x - \text{median}(x)|)}$$

[1]

where $x$ is the distribution of responses. This procedure could be regarded as a non-parametric z-scoring of the data.

The rationale for normalizing our data was twofold. First, we used this procedure to reject outliers in the distribution of responses. Following previous studies (21), we discarded all responses with that deviated from the median in more than 15 times the median absolute deviance. The second purpose of normalization was to average our results across different
questions. This helps the visualization of our data, but our findings can be replicated on each question separately without any normalization (Supplementary Fig. 1).

**Data analysis**

To compute all our curves in Fig. 1, we subsampled our crowd in two different ways: either by choosing \( n \) individuals that interacted in \( m = n/5 \) different groups (within-groups sampling) or by choosing \( n \) individuals from \( n \) different groups (between-groups sampling). All curves in Fig. 1B and the solid line in Fig. 1C were based on the within-groups sampling condition; the dashed line in Fig. 1C is from the between-groups sampling condition. For a fair comparison between conditions, we computed the errors using exactly the same subsamples in our crowd. For each value of \( n \), we considered 1,000 iterations of this subsampling procedure.

In the case of \( n = 5 \), each iteration randomly selected 5 of our 280 complete groups (Fig. 1C sketches one exemplary iteration). In the within-groups condition, we computed the crowd error of each of the 5 groups (the error of the average response in stages \( i_1 \) and \( i_2 \), and the error of the collective response in stage \( c \) ) respecting the identity of each group. Finally, we averaged the 5 crowd errors and stored their mean value as the within-groups error for this iteration. In the between-groups sampling, we combined responses from individuals coming from different groups. We computed the error for 1,000 random combinations contingent on the restriction that all individuals belonged to different groups. Finally, we averaged all crowd errors and stored this value as the between-groups error for this iteration.

The same procedure was extended for \( n > 5 \). We randomly selected \( n \) of our 280 groups on each of our 1,000 iterations. In the within-groups condition, we selected all possible combinations of \( n \) individuals coming from \( m \) groups, and computed their crowd error. We averaged the crowd error for all possible combinations and stored this value as the within-groups error for this iteration. In the between-groups condition, we randomly selected 1,000 combinations of \( n \) individuals coming from \( n \) different groups, and computed their crowd error.
We averaged all of these crowd errors and stored this value as the between-subjects error for this iteration.

All error bars in Fig. 1 depict the normalized mean ± s.e.m. of the crowd error across iterations. Pairwise comparisons were performed through non-parametric paired tests (Wilcoxon sign rank tests). To test the general tendency that error decreases for larger crowds, we used two-way repeated-measured analysis of variance (rm-ANOVA) with factors question and crowd size $n$, and iteration as repeated measure.

References

38. M. Lopez-Rosenfeld et al., Neglect in human communication: quantifying the cost of cell-phone interruptions in face to face dialogs. Plos One. 10, e0125772 (2015).

39. T. Niella, N. Stier-Moses, M. Sigman, Nudging Cooperation in a Crowd Experiment. Plos One. 11, e0147125 (2016).
Supplementary Figure 1. Error as a function of crowd size for each question. Each panel show the same as Fig. 1B for each question separately and without normalization. Blue lines show the error obtained by averaging $n$ individual answers at stage $i1$, red lines show the same for stage $i2$, and black lines show the error of averaging $m$ collective estimates (black lines, stage $c$). All lines show the mean across 1,000 random within-group sub-samples of crowd; the s.e.m. is within the thickness of each line. 

A) GOALS: How many goals were scored in the 2010 FIFA World Cup?

B) ALEGRIA: How many times does the word “alegría” (joy) appear in the lyrics of the song “Y dale alegría a mi corazón”? (“Give joy to my heart”)

C) ROULETTE: What is the sum of all numbers in a roulette wheel?

D) OIL BARREL: How much did a barrel of oil cost in 1970 (in US dollars cents)?
Supplementary Figure 2. Distribution of responses relative to the correct answer. On each question, the top panel is the normalized histogram of responses obtained at stage \( i1 \) (blue solid line). The vertical dashed line is the mean of that distribution. The bottom panel show the distribution of mean estimates on each group (blue dashed line) and the distribution of consensus decisions of each group (black line). If the consensus reflected an average with uniform weights for each participant, then these two distributions should be identical. However, we observe that in the four questions, collective answers are less biased than the mean of all initial estimates of each group (black dashed line is closer to 0 than the vertical dashed line). This is why the wisdom of collective answers is more accurate than the wisdom of \( n=m/5 \) initial estimates. Each panel A-D shows a different question. See Supplementary Table 1 for more details.
| Question                                                                 | Code   | Correct Answer | Mean  | Median | Median Absolute Deviance | Discussed in groups? |
|-------------------------------------------------------------------------|--------|----------------|-------|--------|--------------------------|-----------------------|
| How many goals were scored in the 2010 FIFA World Cup?                  | GOALS  | 145            | 107.9 | 81     | 38                       | Yes                   |
| How many emperors did the Roman Empire have?                           | EMPERORS | 131            | 19.5  | 10     | 5                        | No                    |
| How many times does the word “alegría” (joy) appear in the lyrics of the song “Y dale alegría a mi corazón”? (“Give joy to my heart”) | ALEGRIA | 21             | 26.7  | 21     | 9                        | Yes                   |
| How many calories are there in 200 grams of butter?                     | BUTTER | 1480           | 1251.3| 600    | 400                      | No                    |
| What is the sum of all numbers in a roulette wheel?                    | ROULLETTE | 666            | 705.9 | 400    | 250                      | Yes                   |
| How much did a barrel of oil cost in 1970 (in US dollars cents)?       | OIL BARREL | 180           | 6445.4| 150    | 138                      | Yes                   |
| What is the height of the Eiffel tower in meters?                      | EIFFEL | 324            | 344.4 | 200    | 110                      | No                    |
| How many elevators are there in the Empire State Building of New York? | ELEVATORS | 73             | 18.6  | 11     | 6                        | No                    |

**Supplementary Table 1.** Questions, correct answers, and summary statistics of the distribution of responses before interaction (stage \(i_1\)). From the eight questions asked in stage \(i_1\), half of them were discussed in groups (stage \(c\)). Distributions were normalized using a non-parametric method based on the median (column 5) and the median absolute deviance (column 6). See Equation [1] in Methods for details.