nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-017-0273-4

Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds

Joaquin Navajas[1,2]*, Tamara Niella[1,3], Gerry Garbulsky[4], Bahador Bahrami[2,5] and Mariano Sigman[1]

The aggregation of many independent estimates can outperform the most accurate individual judgement[1-3]. This centenarian finding[1,2], popularly known as the 'wisdom of crowds'[3], has been applied to problems ranging from the diagnosis of cancer[4] to financial forecasting[5]. It is widely believed that social influence undermines collective wisdom by reducing the diversity of opinions within the crowd. Here, we show that if a large crowd is structured in small independent groups, deliberation and social influence within groups improve the crowd's collective accuracy. We asked a live crowd ($N = 5{,}180$) to respond to general-knowledge questions (for example, "What is 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 averaging consensus decisions was substantially more accurate than aggregating the initial independent opinions. Remarkably, combining as few as four consensus choices outperformed the wisdom of thousands of individuals.

Understanding the conditions under which humans benefit from collective decision-making has puzzled mankind since the origin of political thought[6]. Theoretically, aggregating the opinions of many unbiased and independent agents can outperform the best single judgement[1], which is why crowds are sometimes wiser than their individuals[2,3]. This principle has been applied to many problems, including predicting national elections[7], reverse-engineering the smell of molecules[8] and boosting medical diagnoses[4]. The idea of wise crowds, however, is at odds with the pervasiveness of poor collective judgement[9]. Human crowds may fail for two reasons. First, human choices are frequently plagued with numerous systematic biases[10]. Second, opinions in a crowd are rarely independent. Social interactions often cause informational cascades, which correlate opinions, aligning and exaggerating the individual biases[11]. This imitative behaviour may lead to 'herding'[9], a phenomenon thought to be the cause of financial bubbles[12], rich-get-richer dynamics[13,14] and zealotry[15]. Empirical research has shown that even weak social influence can undermine the wisdom of crowds[16], and that collectives are less biased when their individuals resist peer influence[17]. 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 left unexplored the contribution of deliberation in creating wise crowds. For example, allowing individuals to discuss their opinions in an online chat room results in more accurate estimates[18,19]. Even in face-to-face interactions, human groups can communicate their uncertainty and make joint decisions that reflect the reliability of each group member[20,21]. During peer discussion, people also exchange shareable arguments[22,23], which promote the understanding of a problem[24]. Groups can reach consensuses that are outside the span of their individual decisions[24,25], even if a minority[26] or no one[24] 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? We addressed these questions by performing an experiment on a large live crowd (Fig. 1a, see also Supplementary Video 1).

We asked a large crowd ($N = 5{,}180$ (2,468 female), aged 30.1 ± 11.6 yr (mean ± s.d.)) attending a popular event to answer eight questions involving approximate estimates to general knowledge quantities (for example, "What is the height in metres of the Eiffel Tower?" cf. 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 s 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 s to write down their individual estimate, which gave them a chance to revise their opinions and change their minds (stage i2, right panel in Fig. 1a). Participants also reported their confidence in their individual responses on a scale from 0 to 10.

Responses to different questions were distributed differently. To pool the data across questions, we used a non-parametric normalizing method, used for rejecting outliers[27] (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 ($F(4{,}999) = 477.3$, $P \approx 0$; blue lines in Fig. 1b), replicating the classic wisdom-of-crowd effects[2]. The average of all initial opinions in the auditorium ($N = 5{,}180$) led to 52% error reduction compared with the individual estimates (Wilcoxon signed-rank test, $z = 61.79$, $P \approx 0$).

We then focused on the effect of debate on the wisdom of crowds, and studied whether social interaction and peer discussion impaired[16,17] or promoted[23,24] collective wisdom. To disentangle these two main alternative hypotheses, we looked at the consensus

[1]Universidad Torcuato Di Tella, Buenos Aires, Argentina. [2]Institute of Cognitive Neuroscience, University College London, London, UK. [3]Psychology Department, University of Oregon, Eugene, OR, USA. [4]TED, Buenos Aires, Argentina. [5]Faculty of Psychology and Educational Sciences, Ludwig Maximilian University, Munich, Germany. *e-mail: joaquin.navajas@utdt.edu

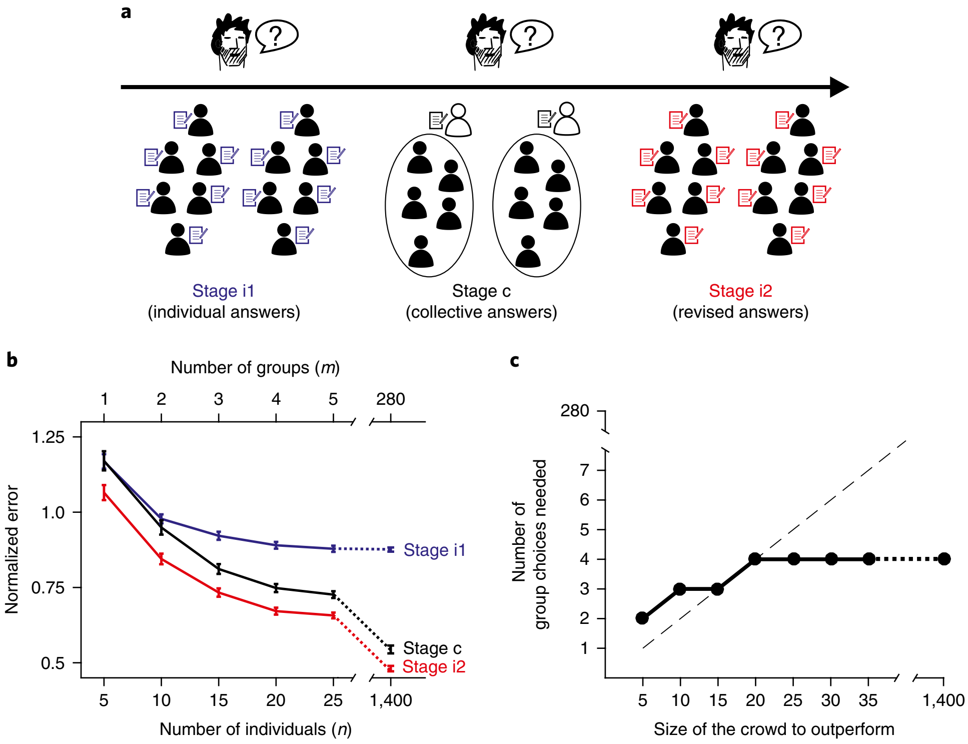

Fig. 1 | Aggregating debates and the wisdom of crowds. a, A live crowd ($N=5{,}180$) answered general knowledge questions in three stages. Left: initial individual estimate (stage i1). Middle: consensus (stage c). Right: revised individual estimate (stage i2). In stage c, a moderator (white) recorded the group's consensus estimate. **b**, Normalized error of the average of n individual answers (blue line for stage i1, red line for stage i2), and normalized error of the average of $m=n/5$ collective estimates (black line, stage c). Error bars are s.e.m. **c**, Minimum number of collective decisions needed to significantly ($\alpha=0.01$) outperform crowds of different sizes. At least two group estimates (from ten individuals) are needed to outperform the wisdom of five independent individuals, and three estimates (from fifteen individuals) are needed to outperform the wisdom of ten independent estimates. For these crowd sizes, the wisdom of crowds is more efficient than aggregating debates. However, averaging four collective decisions leads to estimates that are significantly more accurate than the wisdom of crowds of any size. The thin dashed line shows the number of groups corresponding to each crowd size (five participants per group). When the solid line is below the dashed line, it indicates that averaging the group consensus outperforms averaging individuals.

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, $n=5m$ as there were 5 participants in each group). This analysis is based on the 280 groups (1,400 participants) that had valid data from all of their members (see Methods). We observed that the average of as few as collective estimates was more accurate than the mean of the 15 independent initial estimates (blue line at $n=15$ versus 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 $n=20$ versus black line at $m=4$ in Fig. 1b, $z=20.79$, $P=10^{-96}$). Most notably, the average of 4 collective estimates was even more accurate (by 49.2% reduction in error) than the average of the 1,400 initial individual estimates (blue data point at $n=1{,}400$ versus 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 rule out this possibility as one or two collective estimates were not better than five or ten independent initial estimates, respectively ($z=1.02$, $P=0.31$). In other words, this is the result of a 'crowd of crowds' (Fig. 1c).

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 with initial estimates (blue line versus red line in Fig. 1b, $F(1{,}999)=4{,}458.6$, $P\approx0$). When compared with collective choices, the average of n revised decisions was overall more accurate than the average of m group decisions (black line versus red line in Fig. 1b, $F(1{,}999)=2{,}510.4$, $P\approx0$), although this depended on the specific question asked (interaction $F(3{,}999)=834.7$, $P\approx0$; see Supplementary Fig. 1). Taken together, these findings demonstrate that face-to-face social interaction brings remarkable benefits in accuracy and efficiency to the wisdom of crowds. These results raise the question of how social interaction, which is expected to instigate herding, could have improved collective estimates.

To answer this question, we analysed how the bias and the variance of the distribution of estimates were affected by debates (Fig. 2). Figure 2a shows a graphical representation of how deliberation and social influence affected the distribution of responses in two exemplary groups. We found that the consensus decisions were less biased than the average of initial estimates (Fig. 2b, $z=2.15$, $P=0.03$, see also Supplementary Fig. 2). This indicates that deliberation led to a better consensus than what a simple averaging procedure (with uniform weights) could achieve. 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 (Fig. 2c, $z=27.15$, $P=10^{-162}$). Moreover, in line with previous reports that social influence reduces the diversity of opinions[16,17], 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 (Fig. 2d, Wilcoxon signed-rank test of the variance of

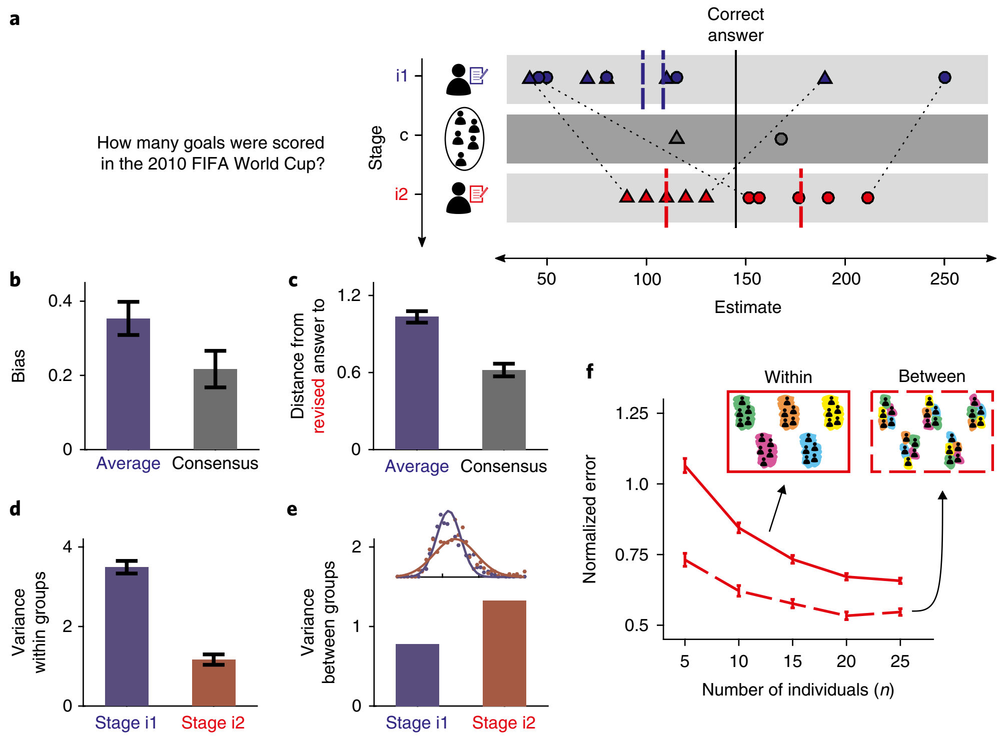

Fig. 2 | The effect of deliberation on bias and variance. a, Schematic of the process of deliberation in two groups (circles or triangles) answering the question GOALS (see Supplementary Table 1) across the three stages of the experiment. Vertical dashed lines show the mean of each group. The range of opinions within each group decreased going from stage i1 to i2 (dotted black lines). The range of the average opinions (distance between vertical dashed lines) increased after deliberation. **b**, Consensus decisions (grey bar) were less biased than the simple average of initial estimates in the group (blue bar). Bars show mean bias (signed error) and s.e.m across $m = 280$ groups. **c**, Revised estimates (obtained in stage i2) were closer to the consensus decision (grey bar) than to the average of initial estimates (blue bar). Bars show mean distance and s.e.m ($n = 1{,}400$). **d**, Individuals conformed to the group consensus. Deliberation decreased the diversity of opinions within groups. Bars show variance within groups (mean ± s.e.m. across $m = 280$ groups) before (blue bar) and after (red bar) deliberation. **e**, Deliberation led to polarization of opinion and pulled groups to wider extremes in the opinion space. This process increased the diversity of opinions between different groups. The between-groups variance is obtained by taking the mean estimate of each group and computing the variance of this distribution across all groups ($m = 280$). The inset shows the distribution of mean estimates before (blue) and after (red) deliberation. Bars show variance between groups. **f**, We aggregated the individual estimates in two different ways: either by sampling participants all from the same 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 colour were averaged together. The *y* axis shows the normalized error of the average of *n* individual answers at stage i2 for the within-groups (solid line) and the between-groups (dashed line) conditions. Sampling participants who interacted in different debates leads to more accurate estimates.

responses on each group before versus after interaction, $z = 18.33$, $P = 10^{-75}$). However, interaction actually increased the variance of responses between groups (Fig. 2e): 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'[28].

Previous studies have proposed that a fundamental condition to elicit the wisdom-of-crowds effect is the diversity of opinions[3,29]. 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. 2e). 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.

Consistent with our predictions, we found that breaking the local correlations by between-group sampling led to a large error reduction (red solid line versus red dashed line in Fig. 2f, 26% error reduction on average, $F(1{,}999) = 25{,}824.1$, $P \approx 0$). In fact, averaging only five revised estimates coming from five different groups outperformed the aggregation of all initial independent decisions in the auditorium ($z = 25.91$, $P = 10^{-148}$). This finding is consistent with previous studies showing that averaging approximately five members of 'select crowds' leads to substantial increases in accuracy[30,31]. In our case, adding more decisions using this sampling procedure led to a significant decrease in error ($F(4{,}999) = 249.34$, $P \approx 0$). 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.

We then asked whether deliberation was necessary to observe an increase in the wisdom of crowds. One could argue that the difference between wisdom of crowds obtained by aggregating the

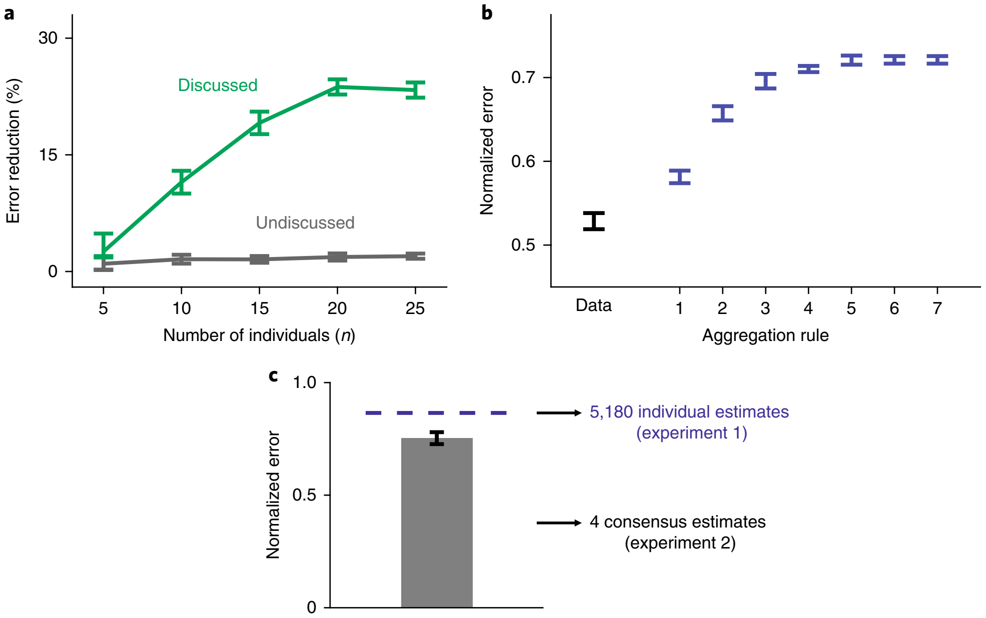

Fig. 3 | The superior wisdom of deliberative crowds. a, Collective accuracy in the absence of deliberation. Error reduction between averaging *n* revised estimates compared with averaging *n* initial estimates, expressed as percentage decrease. The green line shows the mean error reduction (over $N=1,000$ random subsamples) in the discussed questions and the grey line shows the same for the questions that remains undiscussed. Error bars depict s.e.m. **b**, Simple aggregation rules fail to explain the accuracy of deliberative crowds. Normalized error of averaging $N=100$ randomly chosen collective decisions (black bar) versus averaging simulated estimates produced by seven simple aggregation rules (blue bars). Aggregation rules: (1) resistance to social influence, (2) confidence-weighted average, (3) expert rule, (4) median, (5) soft median, (6) mean, (7) robust average. See Methods for details about these rules. Error bars depict s.e.m. **c**, We tested whether four consensus choices could consistently and reliably outperform the wisdom of crowds. Aggregating four random consensus estimates collected in the lab (experiment 2) was more accurate than aggregating all 5,180 individual estimates from the crowd (experiment 1). The *y* axis shows error in normalized units (see Methods) and the error bar depicts s.e.m.

first (i1) versus the second (i2) opinions may have simply resulted from having a second chance to produce an estimate. Indeed, previous research[32–34] has shown consistent improvements drawn from repeatedly considering the same problem in decision-making. To evaluate this possibility, we compared wisdom of crowds obtained from the answers to the discussed versus the undiscussed questions (see Methods). Figure 3a shows the error reduction when comparing the average of *n* revised estimates (i2) with the average of *n* initial estimates (i1), that is, the ratio of red line to blue line in Fig. 1b. We observed that the error reduction in the absence of deliberation (Fig. 3a, grey line) was below 3% for all crowd sizes. With deliberation (Fig. 3a, green line), in contrast, error reduction was significantly larger and increased with increasing number of aggregated opinions ($F(1,999)=3,963.6$, $P\approx 0$, comparing with vesus without deliberation). This result demonstrated that merely having the chance to produce a second estimate was not sufficient, and that deliberation was needed to increase the wisdom of crowds.

While we found that deliberation increased collective accuracy, the results presented so far do not shed light on the specific deliberative procedure implemented by our crowd. In principle, collective estimates could have been the output of a simple aggregation rule different to the mean[17,35]. Alternatively, participants could have used the deliberative stage to share arguments and arrive at a new collective estimate through reasoning[22,23]. This dichotomy between 'aggregating numbers' versus 'sharing reasons' has been discussed in several studies about collective intelligence[36,37]. It has been argued that the normative strategy in predictive tasks is to share and aggregate numbers. Instead, problem-solving contexts require authentic deliberation and sharing of arguments and reasons[37]. Which kind of deliberative procedure did the groups implement in our experiment?

To answer this question, we first compared the accuracy of our consensus estimates with seven different aggregation rules for how to combine the initial estimates (see Methods). Three of these rules were based on the idea of robust averaging[38], namely that groups may underweight outlying estimates (that is, the median rule, the soft median rule and the robust averaging rule; see Methods for details). Three other rules were inspired in previous studies showing that the one individual may dominate the discussion and exert greater influence in the collective decision[21] (that is, the expert rule, the confidence-weighted average rule and the resistance-to-social-influence rule; see Methods for details). As a benchmark, we also compared these rules with the simple average rule. Figure 3b shows the expected error if our crowd implemented each of these rules (blue bars in Fig. 3b). The empirically obtained consensus estimates (black bar in Fig. 3b) were significantly more accurate than all seven aggregation rules ($z>3.99$, $P<10^{-5}$ for all pairwise comparisons between the observed data and all simulated rules). The deliberation procedures implemented by our crowd could not be parsimoniously explained by the application of any of these simple rules.

The above analysis definitively rejects the more simplified models of consensus. But the evidence is not exhaustive and does not necessarily imply positive evidence for the hypothesis that our crowd shared arguments during deliberation. To directly test this hypothesis, we ran a second experiment in the lab (experiment 2, $N=100$, see Methods and Supplementary Fig. 3). Groups of five people first went through the experimental procedure (Fig. 1a). After finishing the experiment, in a debriefing questionnaire, we asked them what deliberation procedure(s) they implemented during the debates. After the end of stage i2 (cf. Fig. 1a), all participants were asked to rate (on a Likert scale from 0 to 10) the extent to which different deliberation procedures contributed to reaching consensus

Table 1 | Deliberation procedures implemented during the debates, as reported by the participants

Deliberation procedure	Question 1	Question 2	Question 3	Question 4
We shared arguments and reasoned together	8.0 ± 0.2 (10)	7.5 ± 0.2 (10)	7.9 ± 0.2 (10)	7.4 ± 0.3 (8)
We followed the individuals who verbally expressed higher confidence during the debate	6.4 ± 0.3 (8)	6.0 ± 0.3 (8)	6.1 ± 0.3 (8)	6.9 ± 0.3 (10)
We discarded the estimates that were most far away from the mean	5.4 ± 0.3 (0)	5.0 ± 0.4 (0)	5.5 ± 0.4 (0)	5.3 ± 0.4 (0)
We followed the individuals who had reported higher confidence in the initial stage	5.2 ± 0.4 (0)	4.6 ± 0.4 (0)	4.6 ± 0.4 (0)	5.1 ± 0.4 (0)
We averaged our estimates	5.0 ± 0.4 (0)	4.4 ± 0.3 (0)	4.0 ± 0.3 (0)	3.5 ± 0.4 (0)
We followed the individuals who were least willing to change their minds	2.3 ± 0.3 (0)	2.4 ± 0.3 (0)	2.1 ± 0.3 (0)	3.3 ± 0.3 (0)

In experiment 2, we asked participants to report the extent to which different deliberation procedures contributed to reaching consensus. Participants used a Likert scale from 0 to 10 (see Methods for details). Values are mean rating ± s.e.m. for each question and procedure. Numbers in brackets are the mode of the distribution of ratings (see also Supplementary Fig. 4). Procedures were sorted by mean rating, but participants rated them in a randomized order. Question 1: GOALS. Question 2: ALEGRIA. Question 3: ROULETTE. Question 4: OIL BARREL (see Supplementary Table 1 for details).

(see Table 1 and Supplementary Fig. 4). The procedure with highest endorsement was "We shared arguments and reasoned together" (mean rating ± s.e.m across all questions: 7.7 ± 0.2, mode rating: 10; $z > 7.05$, $P < 10^{-12}$ for all comparisons, Table 1). We gave participants the opportunity to endorse more than one procedure or even describe a different procedure not appearing in our list. This latter option was selected less than 5% of the time (4.2 ± 2.0%). Overall, our control analyses and new experiment suggest that (1) without deliberation there is no substantial increase in collective accuracy (Fig. 3a), (2) the most salient simple aggregation rules previously proposed in the literature did not explain our findings (Fig. 3b), and (3) participants reported sharing arguments and reasoning together during deliberation (Table 1).

Experiment 2 also allowed us to probe the wisdom of deliberative crowds 'by design'. Since the materials (questions) and procedures were identical between the two experiments, we could formally test whether aggregating the consensus estimates (stage i2) drawn from four groups of five people collected in the lab could predictably and consistently outperform the aggregate of all independent opinions (stage i1) in the crowd. We found that the average of four group estimates collected in experiment 2 was significantly more accurate than the average of all 5,180 initial estimates collected from the crowd ($z = 6.55$, $P < 10^{-11}$, Fig. 3c). It is difficult to overstate the importance of these findings as they call for re-thinking the importance of the deliberation structure in joint decision-making processes. This study opens up clear avenues for optimizing decision processes through reducing the number of required opinions to be aggregated.

Our results are in contrast to an extensive literature on herding[11] and dysfunctional group behaviour[39], 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[40] and peer discussion[24] can increase the understanding of conceptual problems. However, these findings offer a key 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[41], to the aggregation of political and economic forecasts[42] and the design of wiser public policies[43]. Our findings thus provide further support to the idea that combining statistics with behavioural interventions leads to better collective judgements[18]. While our aim was to study a real interacting crowd, face-to-face deliberation may not be needed to observe an increase in collective accuracy. In fact, previous research has shown that social influence in virtual chatrooms could also increase collective intelligence[19,20].

The first study on the wisdom of crowds was regarded as an empirical demonstration that democratic aggregation rules can be trustworthy and efficient[2]. Since then, attempts to increase collective wisdom have been based on the idea that some opinions have more merit than others and set out to find those more accurate opinions by pursuing some ideal non-uniform weighting algorithm[17,31,35]. For example, previous studies proposed to select 'surprisingly popular' minority answers[35] or to average the responses of 'select crowds' defined by higher expertise[31] or by resistance to social influence[17]. 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[44], promoting social inclusion[39], and avoiding elitism or inequality[45,46]. Here, we showed that the wisdom of crowds can be increased by simple face-to-face discussion within groups coupled with between-group sampling. Our simple-yet-powerful idea is that pooling knowledge from individuals who participated in independent debates reduces collective error. Critically, this is achieved without compromising the democratic principle of 'one vote, one value'[47]. This builds on the political notion of deliberative polls as a practical mechanism to solve the conundrum between equality and deliberation. Solving these two things simultaneously is difficult because as more and more people's voices are asked to make a decision, massive deliberation becomes impractical[48,49]. Here, we demonstrated that in questions of general knowledge, where it is easy to judge the correctness of the group choice and in the absence of strategic voting behaviour[50], aggregating consensus choices made in small groups increases the wisdom of crowds. This result supports political theories postulating that authentic deliberation, and not simply voting, can lead to better democratic decisions[51].

Methods

Context. The experiment was performed during a TEDx event in Buenos Aires, Argentina (http://www.tedxriodelaplata.org/) on 24 September 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 behavioural experiments on large TEDx audiences. The first two editions studied the cost of interruptions on human interaction[52], and the use of a competition bias in a 'zero-sum fallacy' game[53].

Materials. Research assistants handed 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 of a behavioural 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 5,180 participants (2,468 female, mean age 30.1 yr, s.d. 11.6 yr) 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). A video of the experiment is available in Supplementary Video 1.

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 (for example, "What is the height in metres 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 example question on the screen, and then read the eight questions. Participants were given 20 s 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 on 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 on page 1 (for example, "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 s. Moderators were given verbal and written instructions to not participate or 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 all of their individual decisions and confidence, including the ones that remained undiscussed. 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 5,180 papers: 4,232 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 are based on those 280 complete groups (1,400 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 Fig. 1). To normalize these distributions, we used a non-parametric approach inspired in the outlier detection literature[27]. We calculated the deviance of each data point x_i around the median, and normalized this value by the median absolute deviance:

$$n_i = \frac{x_i \; - \; \text{median}(x)}{\text{median}(|x \; - \text{median}(x)|)} \tag{1}$$

where x is the distribution of responses. The i represents the ith subject where i goes from 1 to N subjects. 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[27], we discarded all responses that deviated from the median by 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. 2e were based on the within-groups sampling condition; the dashed line in Fig. 2e 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. 2e sketches one example iteration). In the within-groups condition, we computed the crowd error of each of the five groups (the error of the average response in stages i1 and i2, and the error of the collective response in stage c) respecting the identity of each group. Finally, we averaged the five 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 Figs. 1 and 2 depict the normalized mean ± s.e.m. of the crowd error across iterations. Pairwise comparisons were performed through non-parametric paired tests (Wilcoxon signed-rank tests). To test the general tendency that error decreases for larger crowds, we used two-way repeated-measures analysis of variance with the factors 'question' and 'crowd size n', and iteration as repeated measure.

Aggregation rules. We evaluated whether collective estimates could result from seven simple aggregation rules (Fig. 3b). All of these rules predict that the collective estimate j is constructed using a weighted average of the initial estimates x_i with weights w_i:

$$j = \sum_{i=1}^{5} w_i x_i \tag{2}$$

In Fig. 3b, the seven rules were sorted by accuracy. Rule 1 is an average weighted by resistance to social influence. This procedure simulates that, during deliberation, the group follows the individuals who were least willing to change their minds, presumably because they had better information[17]. Resistance to social influence was quantified as the inverse linear absolute distance between the initial (x_i) and revised (r_i) estimates. This quantity was used to compute the weights:

$$w_i = \frac{\sum_j \left| r_j - x_j + \varepsilon \right|^{-1}}{\left| r_i - x_i + \varepsilon \right|} \tag{3}$$

where ε is a constant to prevent divergence when $x_j = r_j$. We simulated this rule using different values of ε ranging from 0.1 to 1,000, and used the value with highest accuracy ($\varepsilon = 1$). In rule 2 (the 'confidence-weighted average rule'), the group uses the initial confidence ratings as weights in the collective decision, $w_i = c_i \,/\, \sum_j c_j$. In rule 3, which we call the 'expert rule', the group selects the estimate of the most confident individual in the group. This rule is defined by $w_i = 1$ for $i = \text{argmax}(\mathbf{c})$, and $w_k = 0$ for $k \neq i$, where $\mathbf{c}$ is a vector with the five initial confidence ratings in the group.

Rule 4 consists of simply taking the median of the initial estimates, which is equivalent to giving a weight $w_i = 1$ to the third-largest estimate in the group, and $w_k = 0$ to all other estimates. Rule 5 is the simple mean, namely $w_i = 0.2$ for all i. Rule 6, which we call 'soft median', is a rule that gives a weight of $w_i = 0.5$ to the third-largest estimate, weights of $w_k = 0.25$ to the second- and fourth-largest estimates, and $w_i = 0$ to the smallest and largest estimates in the group. Finally, rule 7 is a robust average: this rule gives a weight $w_i = 0$ to all estimates in the group that differ from the mean by more than k orders of magnitude, and equal weights to all other estimates. We simulated this rule using different values of k ranging from 1 to 10, and used the value with highest accuracy ($k = 4$).

Experiment 2. A total of $N = 100$ naïve participants (56 female, mean age 19.9 yr, s.d. 1.3 yr) volunteered to participate in our study. Participants were undergraduate students at Universidad Torcuato Di Tella, and were tested as 20 groups of 5. The instructions and procedures were identical to the main task described above. At the end of the experiment, all individuals completed a questionnaire about the deliberation procedure implemented during the task. We asked them to rate (on a Likert scale from 0 to 10) the extent to which different deliberation procedures contributed to reaching consensus for each question. They rated six different procedures (see Table 1 and Supplementary Fig. 4), which appeared in a randomized order. We also gave them the possibility to choose 'other' and describe that procedure.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. The codes that supports the findings of this study are available from the corresponding author upon request.

Data availability. The data that supports the findings of this study are available from the corresponding author upon request.

Received: 20 June 2017; Accepted: 29 November 2017;
Published online: 15 January 2018

References

1. Condorcet, M. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* (L'impremerie royale, Paris, 1785).
2. Galton, F. Vox populi. *Nature* **7**, 450–451 (1907).
3. Surowiecki, J. *The Wisdom of Crowds* (Little, Brown, London, 2004).
4. Kurvers, R. H. et al. Boosting medical diagnostics by pooling independent judgments. *Proc. Natl Acad. Sci. USA* **113**, 8777–8782 (2016).
5. Ray, R. Prediction markets and the financial "wisdom of crowds". *J. Behav. Financ.* **7**, 2–4 (2006).
6. Jowett, B. *The Republic of Plato* (Clarendon Press, Oxford, 1888).
7. Forsythe, R., Nelson, F., Neumann, G. R. & Wright, J. Anatomy of an experimental political stock market. *Am. Econ. Rev.* **82**, 1142–1161 (1992).
8. Keller, A. et al. Predicting human olfactory perception from chemical features of odor molecules. *Science* **355**, 820–826 (2017).
9. MacKay, C. *Extraordinary Popular Delusions the Madness of Crowds* (Wordsworth Editions Limited, Ware, 1841).
10. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and Biases. *Science* **185**, 1124–1131 (1974).
11. Raafat, R. M., Chater, N. & Frith, C. Herding in humans. *Trends Cogn. Sci.* **13**, 420–428 (2009).
12. Chari, V. V. & Kehoe, P. J. Financial crises as herds: overturning the critiques. *J. Econ. Theory* **119**, 128–150 (2004).
13. Salganik, M. J., Dodds, P. S. & Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).
14. Muchnik, L., Aral, S. & Taylor, S. J. Social influence bias: a randomized experiment. *Science* **341**, 647–651 (2013).
15. Festinger, L., Riecken, H. W. & Schachter, S. *When Prophecy Fails: A Social and Psychological Study of a Modern Group that Predicted the End of the World* (Harper-Torchbooks, New York, NY, 1956).
16. Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. How social influence can undermine the wisdom of crowd effect. *Proc. Natl Acad. Sci. USA* **108**, 9020–9025 (2011).
17. Madirolas, G. & de Polavieja, G. G. Improving collective estimations using resistance to social influence. *PLoS Comput. Biol.* **11**, e1004594 (2015).
18. Mellers, B. et al. Psychological strategies for winning a geopolitical forecasting tournament. *Psychol. Sci.* **25**, 1106–1115 (2014).
19. Gürçay, B., Mellers, B. A. & Baron, J. The power of social influence on estimation accuracy. *J. Behav. Decis. Mak.* **28**, 250–261 (2015).
20. Bahrami, B. et al. Optimally interacting minds. *Science* **329**, 1081–1085 (2010).
21. Juni, M. Z. & Eckstein, M. P. Flexible human collective wisdom. *J. Exp. Psychol. Hum. Percept. Peform.* **41**, 1588–1611 (2015).
22. Mercier, H. & Sperber, D. Why do humans reason? Arguments for an argumentative theory. *Behav. Brain. Sci.* **34**, 57–74 (2011).
23. Mercier, H. & Sperber, D. "Two heads are better" stands to reason. *Science* **336**, 979 (2012).
24. Smith, M. K. et al. Why peer discussion improves student performance on in-class concept questions. *Science* **323**, 122–124 (2009).
25. Laughlin, P. R., Bonner, B. L. & Miner, A. G. Groups perform better than the best individuals on letters-to-numbers problems. *Organ. Behav. Hum. Decis. Process.* **88**, 605–620 (2002).
26. Geil, D. M. M. Collaborative reasoning: evidence for collective rationality. *Think. Reason.* **4**, 231–248 (1998).
27. Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **49**, 764–766 (2013).
28. Myers, D. G. & Lamm, H. The group polarization phenomenon. *Psychol. Bull.* **83**, 602–627 (1976).
29. Hong, L. & Page, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl Acad. Sci. USA* **101**, 16385–16389 (2004).
30. Goldstein, D. G., McAfee, R. P. & Suri, S. The wisdom of smaller, smarter crowds. In *Proc. Fifteenth ACM Conference on Economics and Computation Ser.* 471–488 (ACM, Palo Alto, CA, 2014).
31. Mannes, A. E., Soll, J. B. & Larrick, R. P. The wisdom of select crowds. *J. Pers. Soc. Psychol.* **107**, 276–299 (2014).
32. Vul, E. & Pashler, H. Measuring the crowd within: probabilistic representations within individuals. *Psychol. Sci.* **19**, 645–647 (2008).
33. Herzog, S. M. & Hertwig, R. The wisdom of many in one mind: improving individual judgments with dialectical bootstrapping. *Psychol. Sci.* **20**, 231–237 (2009).
34. Ariely, D. et al. The effects of averaging subjective probability estimates between and within judges. *J. Exp. Psychol. Appl.* **6**, 130–146 (2000).
35. Prelec, D., Seung, H. S. & McCoy, J. A solution to the single-question crowd wisdom problem. *Nature* **541**, 532–535 (2017).
36. Lorenz, J., Rauhut, H. & Kittel, B. Majoritarian democracy undermines truth-finding in deliberative committees. *Res. Polit.* **2**, 1–10 (2015).
37. Landemore, H. & Page, S. E. Deliberation and disagreement: problem solving, prediction, and positive dissensus. *J. Pol. Philos. Econ.* **14**, 229–254 (2015).
38. Li, V., Herce Castañón, S., Solomon, J. A., Vandormael, H. & Summerfield, C. Robust averaging protects decisions from noise in neural computations. *PLoS Comput. Biol.* **13**, e1005723 (2017).
39. Asch, S. E. Opinions and social pressure. *Sci. Am.* **193**, 31–35 (1955).
40. Lyman, F. T. in *The Responsive Classroom Discussion: The Inclusion of All Students* (ed. Anderson, A. S.) 113 (Univ. Maryland Press, Potomac, MD, 1981).
41. Dalkey, N. & Helmer, O. An experimental application of the Delphi method to the use of experts. *Manag. Sci.* **9**, 458–467 (1963).
42. Tetlock, P. *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton Univ. Press, Princeton, NJ, 2005).
43. Sunstein, C. R. *Infotopia: How Many Minds Produce Knowledge* (Oxford Univ. Press, Oxford, 2006).
44. Harvey, N. & Fischer, I. Taking advice: accepting help, improving judgment, and sharing responsibility. *Organ. Behav. Hum. Decis. Process.* **70**, 117–133 (1997).
45. Eisenberger, N. I., Lieberman, M. D. & Williams, K. D. Does rejection hurt? An FMRI study of social exclusion. *Science* **302**, 290–292 (2003).
46. Mahmoodi, A. et al. Equality bias impairs collective decision-making across cultures. *Proc. Natl Acad. Sci. USA* **112**, 3835–3840 (2015).
47. Galton, F. One vote, one value. *Nature* **75**, 414 (1907).
48. Mill, J. S. *On Liberty* (John W. Parker and Son, London, 1859).
49. Fishkin, J. S. & Luskin, R. C. Experimenting with a democratic ideal: deliberative polling and public opinion. *Acta Polit.* **40**, 284–298 (2005).
50. Austen-Smith, D. & Banks, J. S. Information aggregation, rationality, and the Condorcet jury theorem. *Am. Political Sci. Rev.* **90**, 34–45 (1996).
51. Cohen, J. in *Deliberative Democracy: Essays on Reason and Politics* (eds Bohman, J. & Rehg, W.) Ch. 3 (MIT Press, Boston, MA, 1997).
52. Lopez-Rosenfeld, M. et al. Neglect in human communication: quantifying the cost of cell-phone interruptions in face to face dialogs. *PLoS ONE* **10**, e0125772 (2015).
53. Niella, T., Stier-Moses, N. & Sigman, M. Nudging cooperation in a crowd experiment. *PLoS ONE* **11**, e0147125 (2016).

Acknowledgements

J.N. and B.B. were supported by the European Research Council StG (NEUROCODEC, #309865); M.S. was supported by the James McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition—Scholar Award (Grant #220020334), and by Agencia Nacional de Promoción Científica y Tecnológica (Argentina)—Préstamo BID PICT (Grant #2013-1653). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank M. Sartorio for assistance in data collection.

Author contributions

J.N., T.N., G.G. and M.S. designed and conducted the experiments. J.N., B.B. and M.S. developed the analysis approach. J.N. analysed the data. J.N., B.B. and M.S. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-017-0273-4.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.N.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Joaquin Navajas

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

▸ Experimental design

1. Sample size

Describe how sample size was determined.	Sample size (N=5180) in Experiment 1 was not pre-determined. This experiment was performed on a live crowd at a public engagement event. Participants were grouped in teams of six individuals (5 players and 1 moderator). Unless explicitly stated, we analyzed data from all groups with complete data for all participants (m=280 groups, n=1400 individuals). The sample size in Experiment 2 was based on the results of Experiment 1 (i.e., that 4 consensus estimates outperformed the wisdom of the entire crowd) and on previous studies (e.g. ref. 17).

2. Data exclusions

Describe any data exclusions.	Given that this experiment was performed in a public engagement event, many participants did not return their answer sheets. This led to incomplete data in many groups, which were excluded from the analysis. No data were excluded from Experiment 2.

3. Replication

Describe whether the experimental findings were reliably reproduced.	In Experiment 1, all experimental findings are based on sub-samples of a live crowd, and were statistically replicated on 1,000 different sub-samples. Experiment 2 replicated the findings of Experiment 1.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.	This experiment did not have a between-subjects design and all participants were allocated to the same experimental group.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.	This experiment did not have a between-subjects design and all participants were allocated to the same experimental group.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	☐	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☒	☐	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

▸ Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

Data were analyzed using Matlab R2016a (Mathworks)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

▸ Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

N/A

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

N/A

b. Describe the method of cell line authentication used.

N/A

c. Report whether the cell lines were tested for mycoplasma contamination.

N/A

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

N/A

▸ Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	In Experiment 1, N=5180, 2468 female, aged 30.1±11.6 years. In Experiment 2, N=100, 56 female, aged 19.9±1.3 years.