Does discussion make crowds any wiser?

Hugo Mercier, Nicolas Claidière

1 DOES DISCUSSION MAKE CROWDS ANY WISER?

2

3 Mercier, H.

4 Institut Jean Nicod,

5 Département d'études cognitives,

6 ENS, EHESS, PSL University, CNRS,

7 Paris France

8 hugo.mercier@gmail.com

9 ORCID ID : 0000-0002-0575-7913

10

11 Claidière, N.

12 Aix Marseille University, CNRS, LPC, FED3C, Marseille, France.

13 nicolas.claidiere@normalesup.org

14 ORCID ID : 0000-0002-4472-6597

15

16 Correspondance to Nicolas Claidière, Laboratoire de Psychologie Cognitive, 3 Place Victor Hugo, 13331

17 Marseille. Nicolas.claidiere@univ-amu.fr

18

19

Abstract

Does discussion in large groups help or hinder the wisdom of crowds? To give rise to the wisdom of crowds, by which large groups can yield surprisingly accurate answers, aggregation mechanisms such as averaging of opinions or majority voting rely on diversity of opinions, and independence between the voters. Discussion tends to reduce diversity and independence. On the other hand, discussion in small groups has been shown to improve the accuracy of individual answers. To test the effects of discussion in large groups, we gave groups of participants (N = 1958 participants in groups of size ranging from 22 to 212; mean 59) one of three types of problems (demonstrative, factual, ethical) to solve, first individually, and then through discussion. For demonstrative (logical or mathematical) problems, discussion improved individual answers, as well as the answer reached through aggregation. For factual problems, discussion improved individual answers, and either improved or had no effect on the answer reached through aggregation. Our results suggest that, for problems which have a correct answer, discussion in large groups does not detract from the effects of the wisdom of crowds, and tends on the contrary to improve on it.

Keywords: Group decision making, wisdom of crowds, aggregation, majority rule, social learning

Word count: 6259 (including references)

Does Discussion Make Crowds Any Wiser?

Ancient Athens is famous for its reliance on democratic decision making. Laws were put forward by a council of 500, and voted by an assembly of 6000 citizens. Judicial decisions were made by courts of 200 jurors (Hansen, 1999). In each case, the assembled citizens would listen to the arguments of the different parties, and the issue would be resolved by a simple majority vote. Crucially, during these votes, discussion among citizens was not formally allowed. Was this a wise rule? If answering this question might have helped Athenians make better decisions, the generalization of democratic decision making means it is an even more pressing question today. Crowds—defined here as any large group, whether or not they are organized—play an increasingly important role, whether in politics—from mass protests to citizens' assemblies—in the creation and diffusion of knowledge—from scientific consortia to Wikipedia contributors—or in business, as companies try to make the best of their workforce's knowledge.

We start by reviewing arguments suggesting that discussion might hinder the wisdom of crowds, and thus that groups might be better off aggregating their answers without discussion, before turning to arguments suggesting instead that discussion might improve the individual performance of the group members, without taking away the added value of the wisdom of crowds. In the absence of empirical evidence directly bearing on this issue, we conduct a large-scale experiment in which 1958 participants in 33 groups with size ranging from 22 to 212 participants (mean 59), are confronted with a variety of problems, first without being able to discuss them, and then with discussion allowed. When an objective benchmark for performance is available, our results suggest that discussion consistently improves individual answers, and also often improves the answer reached through the wisdom of crowds.

In ancient Athens, rules limiting discussion between citizens before a vote were no doubt linked to the practical necessity of making a decision in a limited time frame (often half a day) (Manville & Ober, 2003). More recently, theoretical work has suggested that these constraints might have been wise, maximizing the chances that the citizens would vote for the best available alternative. The most fundamental result underpinning the efficacy of majority voting is the Condorcet Jury Theorem (Condorcet, 1785). For a dichotomous choice, the theorem "states that the probability that a majority votes for the better alternative exceeds *p* [the probability that each voter selects the right option] and approaches 1 as *n* [the number of voters] goes to infinity" (Ladha, 1992, p. 34). The efficacy of majority voting has been demonstrated not only in models (e.g., Austen-Smith & Banks, 1996; Ladha, 1992), but also a variety of experiments (e.g., Hastie & Kameda, 2005).

For the Condorcet Jury Theorem to apply, a set of constraints has to be respected—that the voters are more likely than chance to vote for the best alternative, that they do not vote strategically, and, crucially here, that their decisions are independent of one another. If some voters imitated others, without thinking for themselves, the effective size of the assembly would be reduced, along with the chances that the majority supports the best alternative. During discussion, voters are likely to influence each other, thereby potentially losing some of their independence, and lessening the benefits of majority voting (although see, Estlund, 1994).

Besides majority voting, the other main phenomenon responsible for the wisdom of crowds is averaging. At least since Galton (1907), it has been well established that measures of central tendency such as the mean typically have a lower error than the mean individual error. For instance, when considering a range of numerical estimates that deviate more or less from a correct answer, the error of the mean answer will always be either lower than the mean error (if the correct answer is within the range of all the answers provided), or the same as the

mean error (otherwise) (see, e.g., Larrick & Soll, 2006). Moreover, for many distributions of answers, the error of the mean is uncannily small compared to the mean error, a phenomenon which has allowed averaging to improve performance on a variety of problems ranging from political predictions to medical diagnoses (Surowiecki, 2005).

As in the case of majority voting, the risks of discussion for the benefits of averaging are clear. During discussion, individuals are likely to converge on a middle of the road answer, eliminating the most extreme views, which will reduce the diversity and the range of answers, and lower the potential benefits of averaging. Even increases in individual accuracy might not compensate for this loss of diversity (see, e.g., Hahn et al., 2019; Hong & Page, 2004; Lorenz et al., 2011). There are therefore good grounds to believe that discussion might hamper information aggregation in large groups, which are most likely to benefit from the wisdom of crowds. Indeed, the problem might be particularly acute in the type of densely connected topologies that we will study here (Hahn et al., 2020).

By contrast, other results suggest that discussion might play a positive role. Small-group discussion has been shown to improve the average performance of the group members on a wide range of problems, ranging from logical tasks to political predictions (e.g., Mellers et al., 2014; Moshman & Geil, 1998; Trouche et al., 2014; for reviews, see, Laughlin, 2011; Mercier, 2016; Mercier & Sperber, 2017). In some cases, discussion can even lead to answers that are superior to those reached by any of the group members (e.g., Laughlin et al., 2003). The question remains open of whether this improvement in performance, typically observed in groups of at most five people (although see, Hastie et al., 1983, Hans, 2007 for 12-person juries, with less clearly correct answers, and Mellers et al., 2014 for larger groups interacting through an internet forum), would translate to larger groups, which make discussion less natural (Fay et al., 2000; Krems & Wilkes, 2019), and which might create more opportunities for herding, or for the majority to impose its view regardless of its accuracy (e.g., Asch,

1956).

Still, it is possible that the improvement in performance yielded by small-group discussion might also be observed in larger groups (on the difficulty for accurate answers to spread widely, see, Moussaïd et al., 2017). Improvements in individual performance might then be sufficient to compensate for the decrease in the diversity and independence of the answers, such that discussion will improve, or at least not deteriorate, the wisdom of crowds (be it obtained through majority voting, averaging, or other means of aggregation).

A few studies have tested whether discussion is detrimental to the wisdom of crowds in large groups. In an experiment, mi-sized groups of participants (N = 12) had to make numerical estimates (about, e.g., the population size in a city), and some participants were provided with the average group answer, and an opportunity to revise their estimate on that basis (Lorenz et al., 2011). Although the average performance of these participants improved, several indicators of the strength of the wisdom of crowds decreased (e.g. the degree of diversity within the answers). Another study confirmed that receiving the average answer from other participants leads to a decrease in diversity, but it also found that, for some network configurations, the increase in individual accuracy more than compensated for this loss of diversity (Becker et al., 2017). Importantly, in this latter experiment the participants received the average group answer, but they were expressively forbidden from discussing with one another. Several studies have shown that the increases in accuracy following discussion are substantially larger than those following mere exposure to others' opinion (e.g., Liberman et al., 2012; Minson et al., 2011). This experiment might thus underestimate the benefits of discussion.

In another experiment, a very large crowd (N = 5180) also had to provide numerical estimates of various quantities (Navajas et al., 2018). Crowd members were then provided with the opportunity to talk to each other in small groups (N = 5), for a very short amount of

time (1 min), and to revise their initial answers on the basis of this discussion. In this case, discussion had an unambiguously positive effect, as it increased not only individual performance, but also the answer reached through the wisdom of crowds. However, this study relied on the well-established improvement in performance following small-group discussion, and does not directly address the question of whether a broader discussion within the crowd would also yield such positive effects.

To the best of our knowledge, the study that most directly tested the effect of discussion in medium sized groups (N = 11 to 25) used the following method—which we describe in greater details, since it is similar to the method of the present experiments (Claidière et al., 2017). In each group, participants were seated together in a room, following a grid pattern. The participants were shown a logical or mathematical problem to solve, and given five minutes to attempt to find an answer on their own. Participants then either had fifteen minutes to talk about the problem with their neighbors (Discuss Condition), or to see the response of their neighbors, without discussion (Silence Condition). Every minute, participants recorded their answers, which allowed measuring changes in the percentage of correct answers with time. After the initial five minutes of solitary reasoning, performance improved faster in the Discuss than in the Silence condition. Moreover, a reanalysis of these data shows that discussion vastly improved on the ability of the wisdom of crowds (here, majority voting) to select the best answer. At the end of the first phase of solitary reasoning, the correct answer was supported by the majority of the participants in only 3 out of 12 groups, while it was supported by the majority in all groups after discussion.

Even if this latter study shows that discussion improve individual answers and the aggregated answer yielded by the wisdom of crowds, it has several limitations. The group size, while larger than that used in most experiments on group decision making, was still modest. The problems used were known to yield massive improvement with small-group

discussion (Trouche et al., 2014). The participants were a homogenous group of students. Finally, only one method of aggregating opinions—majority voting—was tested. A measure of central tendency, for instance, might be more sensitive to a loss of diversity following discussion (Hong & Page, 2004; Lorenz et al., 2011).

This overview of the literature suggests that there is no clear existing answer to the question of whether large groups are better off discussing before their opinions are aggregated. To start answering this question, we took advantage of a science festival, the *European Researchers' Night* which would be attended by hundreds of people across 11 towns in France. In each town, a room was set up in which participants could take part in the present experiment, as an introduction to research. As in the Discuss condition of Claidière *et al.* (2017), after being presented with a problem, participants had five minutes to think about it on their own, before being able to discuss it with their immediate neighbors for 15 minutes, with their answers being recorded every minute.

We used three types of problems. First, two *demonstrative problems*, one of which being the bat and ball from the Cognitive Reflection Test (Frederick, 2005). Demonstrative problems have a solution that can be conclusively demonstrated using shared knowledge (Laughlin & Ellis, 1986). These problems constitute an extension and a replication to large, more diverse groups, of the experiment described above (Claidière et al., 2017).

Second, we used two *factual problems*, drawn from Navajas *et al.* (2018), such as "How many goals were scored in the XXX world cup?" If small-group discussion has been shown to improve performance on such problems (Navajas et al., 2018; Sniezek & Henry, 1989), the effects of large-group discussion, and the repercussions of the discussion for the value of the wisdom of crowds, have not been established to the best of our knowledge.

Finally, we used two *ethical problems*, drawn from (Thorndike, 1937), such as "How much money should be awarded to compensate someone who lost a little finger in a

workplace accident?" Discussion in small groups on such problems typically does not lead to systematic changes of mind (Mercier et al., 2017). We did not expect that large-group discussion would lead to different outcomes. As a result, these problems were used as a control in which we did not expect discussion to have any systematic effect on the answers.

If we expect the effects of small-group discussion to also be observed in large groups (as in Claidière et al., 2017 for demonstrative problems), we can derive the following hypotheses:

H1a For demonstrative problems, discussion improves performance more than solitary thinking.

H1b For factual problems, discussion improves performance more than solitary thinking.

H1c For ethical problems, discussion does not have a larger impact than solitary thinking.

When it comes to demonstrative problems, previous results also lead to the prediction that discussion will improve both individual performance and the aggregate answer.

H2 For demonstrative problems, discussion leads to better aggregate answers, as selected through majority voting.

By contrast, for factual problems, it is unknown whether the loss of diversity and independence will compensate for any potential individual gain in accuracy. As a result, we formulate the following research question:

RQ1 For factual problems, does discussion lead to worse, equivalent, or better aggregate answers, as selected through averaging?

Method

Participants

The experiment was part of the *European Researchers' Night*, a pan-European science fair organized by researchers to introduce the public to the world of science and research. In France, the organizing committee of the 2017 edition gave us the opportunity to organize a large participative experiment involving 11 cities and 1958 participants (1048 females). Participants were visitors to the science fair, who came in a large room to take part in an experiment advertised as not being suitable for children younger than 12 (90% of participants reported an age between 13 and 60; median = 24). There were two to six consecutive groups in each city (totaling 33 groups ranging from 20 to 208 individuals [mean 58]), which led to a total of between four to seven groups (259 to 468 participants) per problem. More details can be found in the ESM.

Materials

The six problems we used as material were:

Paul and Linda (demonstrative problem 1). Paul looks at Linda; Linda looks at John; Paul is married; John isn't married; Is someone married looking at someone who isn't married? *Answers provided*: Yes [correct] / No / We can't tell.

Bat and Ball (demonstrative problem 2). A candy and a baguette cost 1.10€ together. The baguette costs 1€ more than the candy. How much does the candy cost? *Correct answer:*

0.05€.

World Cup (factual problem 1). How many goals were scored in the football world cup of 2010? *Correct answer:* 145

Elevators (factual problem 2). How many elevators are there in New York's Empire State Building? *Correct answer:* 73

Little Finger (ethical problem 1). How much money should be awarded to compensate someone who lost a little finger in a workplace accident?

Worms (ethical problem 2). How much money should be awarded to compensate someone who finds they have been eating earthworms in their restaurant meal?

Procedure

The experiment took part in large rooms with chairs arranged in a grid pattern. As participants arrived, they were asked to sit close to each other so that their seating arrangement would be as close as possible to a square grid, with no empty seats. Once everyone was seated, a trained researcher explained to the participants that they were taking part in a real experiment, that they could leave the room at any time, that their anonymous data would be used in a scientific publication and that by giving us their response sheet at the end of the experiment they agreed to these conditions.

Answer sheets were distributed that contained 15 rows, one row for each time step, with the space for an answer to the problem, some demographic questions that were answered immediately (group number, seat number, town, age, gender), and a white space for free

writing. After a brief explanation of the Silence Phase of the experiment, and the importance of not talking, showing each other their answers, or using their phones to check the answer, the experiment started. The problem was displayed on a large screen so that all participants could start answering it at the same time. After 20s, the participants provided their first answer. Four more answers were gathered at succeeding 1-min intervals.

Participants were then told that they would now be able to discuss their answers with their neighbors (Discussion Phase). Neighbors were defined as the eight (maximum) participants surrounding them. Participants were told that the goal was for them to reach a consensus. After they were given the signal to start discussing, the participants had to write down their answer every minute, as in the Silence Phase, for 10 minutes. Time was kept by the experimenter who prompted everyone to write down their answer every minute. At the end of the experiment a 15 min debrief explained the state of the art in group decision making, the purpose of the experiment, and the hypotheses. Participants were also encouraged to advertise the experiment to other potential participants at the fair, but without revealing its purpose and proceedings. Importantly, we changed problems between the groups in each city in order to make sure that participants were completely naïve (i.e. even if they had been informed by a previous participant, they would face a different problem).

Data coding and analysis

Response sheet for demonstrative and factual problems were coded using a crowdsourcing platform. Three independent coders coded the responses of each participant and when available the modal response was retained. In cases in which three different coders disagreed, often due to mistyping from the coders, the experimenters returned to the original response sheet to determine the most likely response (less than 1% of the responses were reevaluated).

Regarding ethical problems, that required more judgment, one independent coder coded all responses from the 499 participants using four categories (for Little finger: a number, a monthly allowance, cost of medical intervention and other; for Worms: a number, the price of the meal, medical costs, and other).

Data exclusion and response variable

We excluded a total of 11% of responses from analysis. This percentage varied between problems, but, crucially, it did not vary with time (see ESM for detailed table). For Paul and Linda, we excluded responses that were not any of the three proposed options (<1%) and used as response variable a binary variable with 1 for correct response and 0 for any of the other two responses. For Bat and Ball, we excluded responses that were not 5 or 10 cents (6%) and used a similar binary variable, with 1 for correct response and 0 for the incorrect response. For the Elevators and World Cup problems we excluded responses that were not numeric, and responses above the 99% quantile to avoid extremely large values (such as "123456"; 7% and 12% of data were excluded resp.).

Finally, for the Worms and Little Finger problems, we excluded data from the "other" category (25% and 27% resp.) and re-coded responses as a binary response variable with 1 being the most frequent response at the end of the Silence Phase (i.e. the majority option before discussion) and 0 for all alternative responses. We should note, however, that our ethical problems, which had no correct answer, were quite different from the other problems and raised a number of issues, such that no strong conclusion can be drawn from them. Based on the advice of reviewers we decided to present the results of the ethics problems in the ESM only.

Statistical method

Analysis were carried out using R (R Core Team, 2020), mixed models were analyzed with the package lme4 (Bates et al., 2015) and ggplot2 was used to produce the figures (Wickham, 2016).

Data availability

All the data analyzed here are available at DOI: 10.17605/OSF.IO/CFWV2

Results

To test H1a, b, and c, we sought to determine whether discussion had a larger effect on the answers than solitary reflection. Figure 1 summarizes the evolution through time of the different groups with the average response for each problem (Supplementary videos 1 to 6 illustrate this evolution using the spatial layout of the rooms in which the experiment was carried out; the videos of each group are available in the public repository of the experiment). Following Claidière et al. (2017), we used mixed models to study the interaction between the experimental phase (Silence vs. Discussion), and time during the first 10 timesteps (to maintain the same number of observations in the two phases: 5 in each of the Silence and Discussion phases). We report the models that combined the problems of each type; however, we also analyzed each problem independently and found that the results of the combined models also applied to each problem independently (full reporting of the models can be found in the Electronic Supplementary Materials). As in our previous study we found that discussion favored the dissemination of the correct response for the two demonstrative problems (β = 0.38, SE = 0.04, z = 8.37, $p < 0.001$). For the two factual problems, there was also a significant interaction between the Silence and Discussion phases, with a reduction in the distance to the correct response observed only during the Discussion Phase (β = -2.31, SE = 0.74, df = 6586, t = -3.12, p = 0.002; see Fig.2).

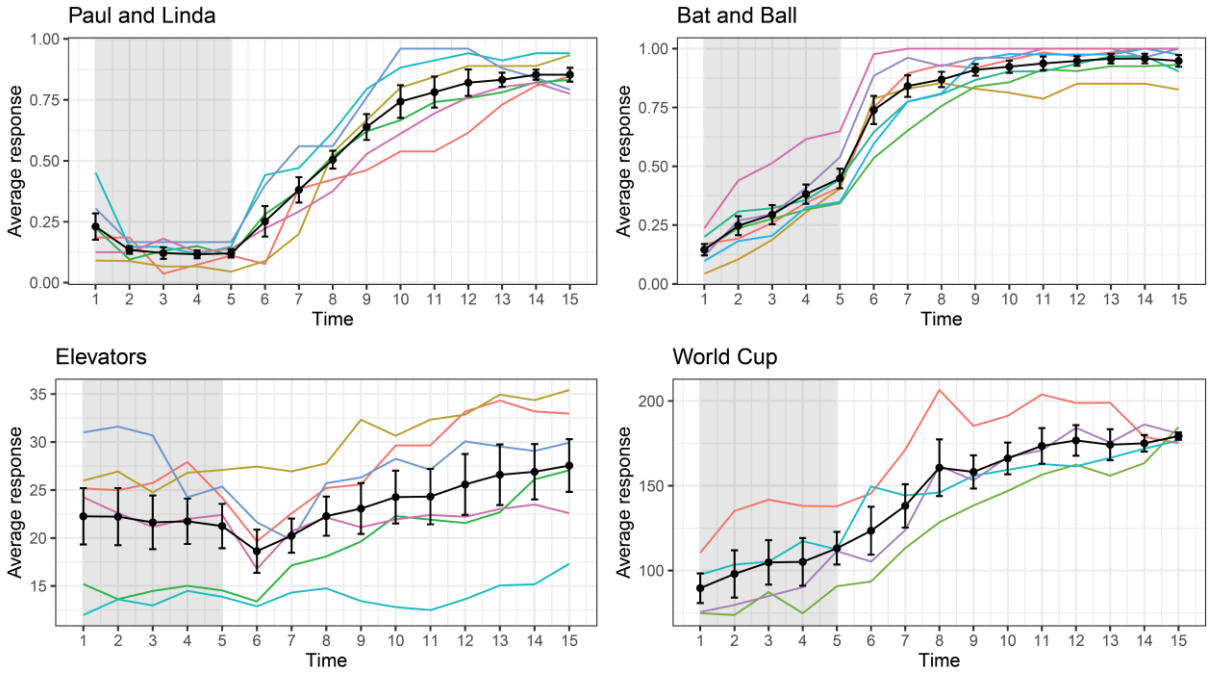

Figure 1: Evolution of the group response for each problem through the Silence (shaded area) and Discussion phases. Each colored line represents a unique group mean response and the black line represents the between group mean (+/- SE). The correct answer to the Elevators problem was 73 and to World Cup problem 145.

To test H2, and answer RQ1, we turn to the effect of discussion on the aggregate answers. For demonstrative problems, we find that discussion leads to better aggregate answers. At the end of the Silence Phase, out of 13 groups, only two had a majority of correct responses (both for the Bat and Ball). By contrast, all groups had a majority of correct responses at the end of the Discussion Phase (a significant improvement, McNemar's chi-squared = 9.10, df = 1, p = 0.003).

For factual problems we found that the error of the mean response (how the mean response in each group differed from the correct answer) decreased for the Elevators problem (Fig. 2; all six groups had a lower error of mean at the end of the Discussion Phase compared to the end of the Silence Phase; binomial test, p = 0.03). By contrast, there was no evidence of a decrease for the World Cup problem (two groups had a value that increased and two a value

that decreased). A possible cause of this difference between the two problems is discussed below.

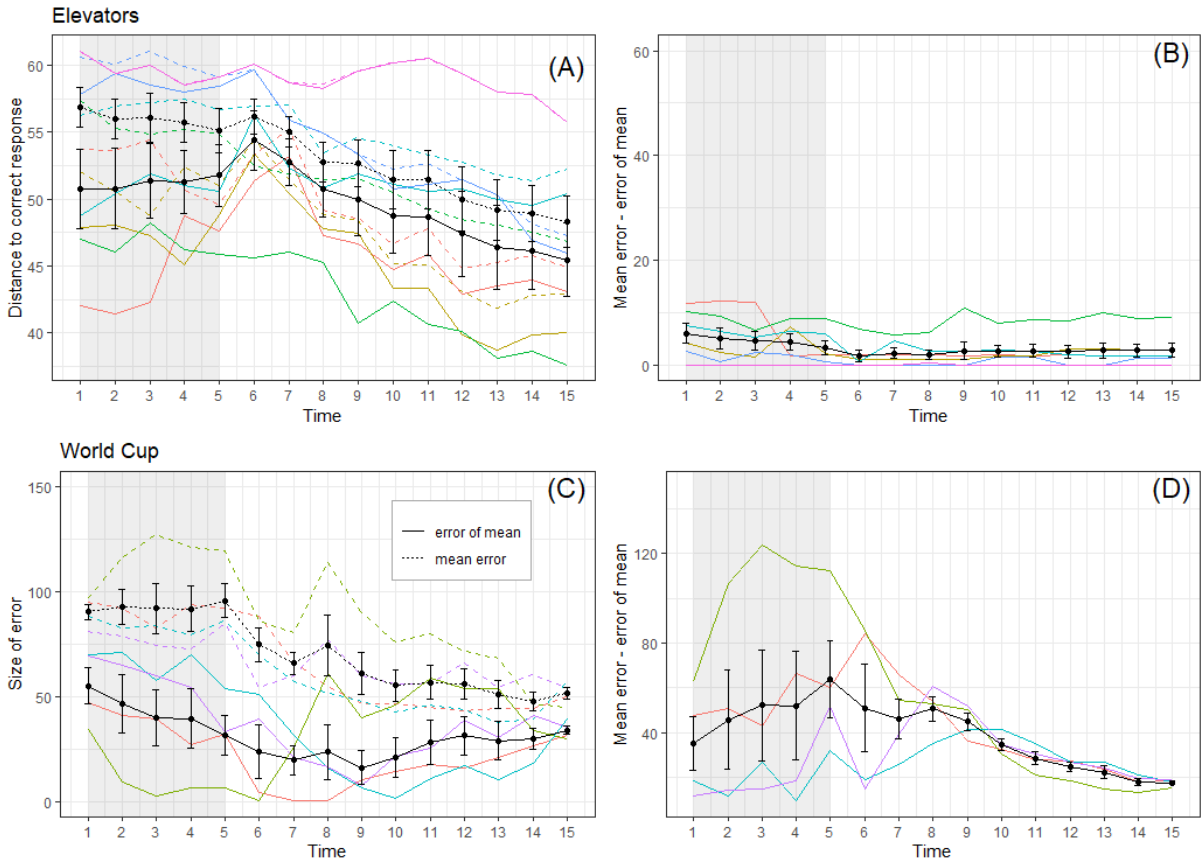

Figure 2: Effect of discussion on the wisdom of crowds. Evolution of the mean error made by individuals (the mean of all the individual errors) and the error of the mean response (mean responses which are depicted in Figure 1) (A, C), as well as the difference between the two (B, D) for the Elevators and World Cup problems. Each colored line represents a unique group mean response and the black line represents the between group mean (+/- SE).

Discussion

H1a, b, and c were confirmed. For both demonstrative and factual problems, discussion improved performance over solitary thinking. The results also clearly supported H2: for demonstrative problems, discussion improved not only individual answers, but also the answers favored by majority voting, which went from two correct answers at the end of beginning of the Discussion Phase, to 13 out of 13 at the end.

Regarding RQ1, the answer is more equivocal. For one factual problem (Elevators),

discussion consistently improved not only on individual answers, but also on the answers reached through averaging within each group. By contrast, for the other factual problem (World Cup), discussion improved on individual answers, but not on the answers reached through averaging.

To understand the differential impact of discussion on the wisdom of crowds in the two factual problems, it is useful to go back to Figure 2. As noted previously, the mean error of the participants decreased through time for both problems. Moreover, the wisdom of crowds effect was present throughout the experiment, with the error of the mean being always inferior to the mean error of individuals (Fig. 2A, 2C). However, while the size of the gain through aggregation (i.e. the difference between the mean error and the error of the mean) stayed relatively constant during the Discussion Phase for Elevators (Fig. 2B), it decreased for World Cup (Fig. 2D).

To make sense of this difference, we can consider two ways for the mean error to decrease: (i) if most answers are distant from the correct answer, and there is a directional shift towards the correct answer, or (ii) if most answers aren't too distant from the correct answer, and there is a reduction of the variance in the answers, with the most extreme answers converging towards the correct answer. Overall, in Elevators, there is no decrease in variance (Fig. 2A), but there is a general shift towards the correct answer, which the overwhelming majority of participants had initially underestimated (Fig. 1). By contrast, in World Cup, there is no directional shift towards the correct answer, with the average answer being as distant from the correct answer at the beginning than at the end of the discussion (Fig 1); however, there is a reduction in the variance of the answers (Fig. 2C). Such a reduction in variance lowers the mean error, but not the error of the mean, thereby decreasing the difference between the two.

It is also worth noting that in all but one of the 10 groups facing factual problems, on

average participants moved more towards the correct answer than towards what was the average group answer at the beginning of the discussion (see ESM, Table S3, and Fig. S2). Indeed, on the whole participants barely moved towards the average answer (Elevators, 1.34; World Cup, 0.10), but they consistently moved towards the correct answer (Elevators, 7.30; World Cup, 36.03). This means that the improvement observed during discussion did not result from participants simply converging towards an answer corresponding to the average at the beginning of the Discussion Phase, as might be expected if participants felt the pull of the majority (see, e.g., Moussaïd et al., 2013). Instead, in every group participants moved towards the correct answer. For factual problems (as for logical problems), in the course of discussion participants appear to have been pulled by arguments towards the correct answer (see, Claidière et al., 2017; Mercier & Sperber, 2017).

Conclusion

Are crowds wiser with or without discussion? The literature makes conflicting predictions, and to answer this question we gave groups of medium to large size (N = 20 to 208) a problem to tackle individually first, and then through discussion with their neighbors. When there were objective benchmarks, individual answers consistently improved with discussion, while aggregate answers improved in most cases and never consistently worsened.

When it comes to problems for which a correct answer exists, our results strongly argue in favor of discussion. First, for the four problems with correct answers studied here—two logical, demonstrative problems, and two factual problems—discussion always improved the mean individual answer. Second, in three out of four cases, discussion led to better aggregate answers, aggregated either through majority voting (the two demonstrative problems), or through averaging (one factual problem). Third, in the last case with no improvement in aggregate answers, discussion was not detrimental to the aggregated answer because it had no effect. Thus, discussion had no detrimental effect on the wisdom of crowds

for the problems examined here.

Our results also demonstrate the effectiveness of discussion in a more qualitative manner. For the two demonstrative problems, 15 minutes of discussion yielded enormous improvements in individual answers, which moved from 12% correct to 84% correct for Paul and Linda, and from 41% correct to 91% correct for the Bat and Ball. Remarkably, in the case of Paul and Linda, all groups reached at least 75% of correct answers, even though they had started with at best 17%. These results thus demonstrate the robustness of the 'truth-wins' scheme, by which a single individual with a correct answer to a demonstrative problem can convince a group, since we also observe its effects in large and diverse groups.

The positive effects of discussion are also clear for the two factual problems. In the Elevator problem, all groups correctly increased their average answer through discussion, moving from a mean error of 55 at the beginning of the discussion to a mean error of 48 at the end. In the World Cup problem, discussion nearly halved the mean error from 96 to 52. We also note that asking participants to estimate the number of goals scored in one specific world cup is a very high bar and it is remarkable that the average number of goals scored in the past six world cups is 160 goals, a difference of only 19 goals with the grand average reached at the end of the discussion. Moreover, in our experiments, participants were constrained in terms of who they could discuss the problems with. Giving people flexibility in network formation might further increase the advantages of discussion (see, e.g., Almaatouq et al., 2020). Alternatively, constraining networks to optimize the flow of information has also been shown to improve accuracy when discussion is not possible, but the same results might extend to situation in which discussion is possible (Jönsson et al., 2015).

Our results have theoretical and practical consequences. They support theoretical frameworks that postulate the power of discussion to change minds for the best (Mercier & Sperber, 2011, 2017), and they show that the loss in independence and diversity in the

answers during discussion can be largely compensated by the increase in accuracy, contrary to what had been suggested (e.g., Hong & Page, 2004; Lorenz et al., 2011). Practically, our results show that discussion is a robust tool to improve not only individual, but also collective answers, even in large and diverse groups, at least for problems that have a correct answer.

Acknowledgments

We thank all the staff of the French *European Researchers' Night*, the local and the national organizing committees and in particular Matteo Merzagora and Lionel Maillot for their help in setting up the "Grande Experience Participative". We also thank all the people that conducted the experiment, Sacha Altay, François Druelle, Justine Epinat, Annabelle Goujon, Julie Gullstrand, Sébastien Lerique, Heather McLeod, Mathilde Menoret, Virginie Postal-Le Dorse, Jean-Pierre Thibaut, Romain Trincherini, and Jean-Baptiste Van der Henst. The authors declare no competing interests. NC gratefully acknowledges financial support from ASCE (ANR-13-PDOC-0004) and LICORNES (ANR-12-CULT-0002). HM gratefully acknowledges financial support from FrontCog (ANR-17-EURE-0017), and PSL (ANR-10-IDEX-0001-02).

References

Almaatouq, A., Noriega-Campero, A., Alotaibi, A., Krafft, P. M., Moussaid, M., & Pentland, A. (2020). Adaptive social networks promote the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *117*(21), 11379–11386.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs*, *70*(9), 1–70.

Austen-Smith, D., & Banks, J. S. (1996). Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review*, *90*(01), 34–45.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences*, *114*(26), E5070–E5076.

Claidière, N., Trouche, E., & Mercier, H. (2017). Argumentation and the diffusion of counter-intuitive beliefs. *Journal of Experimental Psychology: General*, *146*(7), 1052–1066.

Condorcet. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'imprimerie royale.

Estlund, D. (1994). Opinion leaders, independence, and Condorcet's jury theorem. *Theory and Decision*, *36*(2), 131–162.

Fay, N., Garrod, S., & Carletta, J. (2000). Group discussion as interactive dialogue or as serial monologue: The influence of group size. *Psychological Science*, *11*(6), 481–486.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

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

Hahn, U., Hansen, J. U., & Olsson, E. J. (2020). Truth tracking performance of social networks: How connectivity and clustering can make groups less competent. *Synthese*, *197*(4), 1511–1541.

Hahn, U., von Sydow, M., & Merdes, C. (2019). How communication can make voters choose less well. *Topics in Cognitive Science*, *11*(1), 194–206.

Hans, V. P. (2007). Deliberation and dissent: 12 angry men versus the empirical reality of juries. *Chi.-Kent L. Rev.*, *82*, 579.

Hansen, M. H. (1999). *The Athenian democracy in the age of Demosthenes: Structure, principles, and ideology*. University of Oklahoma Press.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*(2), 494–50814.

Hastie, R., Penrod, S., & Pennington, N. (1983). *Inside the Jury*. Harvard University Press.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(46), 16385.

Jönsson, M. L., Hahn, U., & Olsson, E. J. (2015). The kind of group you want to belong to: Effects of group structure on group accuracy. *Cognition*, *142*, 191–204.

Krems, J. A., & Wilkes, J. (2019). Why are conversations limited to about four people? A theoretical exploration of the conversation size constraint. *Evolution and Human Behavior*, *40*(2), 140–147.

Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 617–634.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*, 111–127.

Laughlin, P. R. (2011). *Group problem solving*. Princeton University Press.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology, 22*, 177–189.
Laughlin, P. R., Zander, M. L., Knievel, E. M., & Tan, T. S. (2003). Groups perform better than the best individuals on letters-to-numbers problems: Informative equations and effective reasoning. *Journal of Personality and Social Psychology, 85*, 684–694.
Liberman, V., Minson, J. A., Bryan, C. J., & Ross, L. (2012). Naïve realism and capturing the "wisdom of dyads." *Journal of Experimental Social Psychology*, *48*(2), 507–512.
Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025.
Manville, B., & Ober, J. (2003). *A Company of Citizens: What the World's First Democracy Teaches Leaders About Creating Great Organizations*. Harvard Business Review Press.
Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., & others. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115.
Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in Cognitive Sciences*, *20*(9), 689–700.
Mercier, H., Castelain, T., Hamid, N., & Marín Picado, B. (2017). The power of moral arguments. In J. F. Bonnefon & B. Trémolière (Eds.), *Moral Inferences*. Psychology Press.
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57–74.
Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press.
Minson, J. A., Liberman, V., & Ross, L. (2011). Two to Tango. *Personality and Social Psychology Bulletin*, *37*(10), 1325–1338.
Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking and Reasoning*, *4*(3), 231–248.
Moussaïd, M., Herzog, S. M., Kämmer, J. E., & Hertwig, R. (2017). Reach and speed of judgment propagation in the laboratory. *Proceedings of the National Academy of Sciences*, *114*(16), 4117–4122.
Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013). Social influence and the collective dynamics of opinion formation. *PloS One*, *8*(11), e78433.
Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, *2*(2), 126.
R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, *43*(1), 1–28.
Surowiecki, J. (2005). *The Wisdom Of Crowds*. Anchor Books.
Thorndike, E. L. (1937). Valuations of certain pains, deprivations, and frustrations. *The Pedagogical Seminary and Journal of Genetic Psychology*, *51*(2), 227–239.
Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971.
Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.