



**HAL**  
open science

## Does discussion make crowds any wiser?

Hugo Mercier, Nicolas Claidière

► **To cite this version:**

Hugo Mercier, Nicolas Claidière. Does discussion make crowds any wiser?. *Cognition*, 2021, pp.104912. 10.1016/j.cognition.2021.104912 . hal-03447665

**HAL Id: hal-03447665**

**<https://hal.science/hal-03447665>**

Submitted on 24 Nov 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# DOES DISCUSSION MAKE CROWDS ANY WISER?

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](mailto: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](mailto: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](mailto:Nicolas.claidiere@univ-amu.fr)

18

19

## DOES DISCUSSION MAKE CROWDS ANY WISER?

20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40

### **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?

41 Does Discussion Make Crowds Any Wiser?

42

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

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

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

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

95         As in the case of majority voting, the risks of discussion for the benefits of averaging  
96 are clear. During discussion, individuals are likely to converge on a middle of the road  
97 answer, eliminating the most extreme views, which will reduce the diversity and the range of  
98 answers, and lower the potential benefits of averaging. Even increases in individual accuracy  
99 might not compensate for this loss of diversity (see, e.g., Hahn et al., 2019; Hong & Page,  
100 2004; Lorenz et al., 2011). There are therefore good grounds to believe that discussion might  
101 hamper information aggregation in large groups, which are most likely to benefit from the  
102 wisdom of crowds. Indeed, the problem might be particularly acute in the type of densely  
103 connected topologies that we will study here (Hahn et al., 2020).

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

116 1956).

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

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

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

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

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

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

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

184 Second, we used two *factual problems*, drawn from Navajas *et al.* (2018), such as  
185 “How many goals were scored in the XXX world cup?” If small-group discussion has been  
186 shown to improve performance on such problems (Navajas et al., 2018; Sniezek & Henry,  
187 1989), the effects of large-group discussion, and the repercussions of the discussion for the  
188 value of the wisdom of crowds, have not been established to the best of our knowledge.

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

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

198

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

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

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

205

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

208

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

211

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

215

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

218

219 **Method**

### 220 **Participants**

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

231

### 232 **Materials**

233 The six problems we used as material were:

234

235 *Paul and Linda* (demonstrative problem 1). Paul looks at Linda; Linda looks at John; Paul is  
236 married; John isn't married; Is someone married looking at someone who isn't married?

237 *Answers provided:* Yes [correct] / No / We can't tell.

238

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

241 0.05€.

242

243 *World Cup* (factual problem 1). How many goals were scored in the football world cup of

244 2010? *Correct answer: 145*

245

246 *Elevators* (factual problem 2). How many elevators are there in New York's Empire State

247 Building? *Correct answer: 73*

248

249 *Little Finger* (ethical problem 1). How much money should be awarded to compensate

250 someone who lost a little finger in a workplace accident?

251

252 *Worms* (ethical problem 2). How much money should be awarded to compensate someone

253 who finds they have been eating earthworms in their restaurant meal?

254

### 255 **Procedure**

256 The experiment took part in large rooms with chairs arranged in a grid pattern. As

257 participants arrived, they were asked to sit close to each other so that their seating

258 arrangement would be as close as possible to a square grid, with no empty seats. Once

259 everyone was seated, a trained researcher explained to the participants that they were taking

260 part in a real experiment, that they could leave the room at any time, that their anonymous

261 data would be used in a scientific publication and that by giving us their response sheet at the

262 end of the experiment they agreed to these conditions.

263 Answer sheets were distributed that contained 15 rows, one row for each time step,

264 with the space for an answer to the problem, some demographic questions that were answered

265 immediately (group number, seat number, town, age, gender), and a white space for free

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

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

283

### 284 **Data coding and analysis**

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

295

### 296 *Data exclusion and response variable*

297           We excluded a total of 11% of responses from analysis. This percentage varied  
298 between problems, but, crucially, it did not vary with time (see ESM for detailed table). For  
299 Paul and Linda, we excluded responses that were not any of the three proposed options (<1%)  
300 and used as response variable a binary variable with 1 for correct response and 0 for any of  
301 the other two responses. For Bat and Ball, we excluded responses that were not 5 or 10 cents  
302 (6%) and used a similar binary variable, with 1 for correct response and 0 for the incorrect  
303 response. For the Elevators and World Cup problems we excluded responses that were not  
304 numeric, and responses above the 99% quantile to avoid extremely large values (such as  
305 “123456”; 7% and 12% of data were excluded resp.).

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

314

### 315 *Statistical method*

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

319

### 320 **Data availability**

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

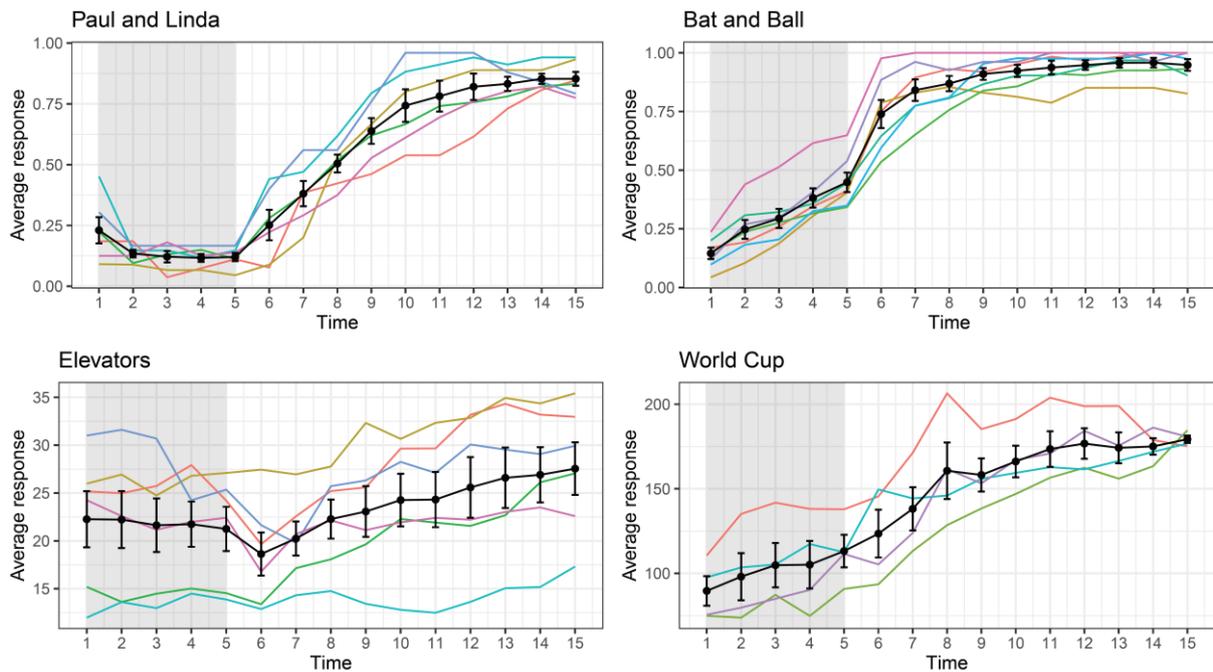
322

323

### **Results**

324           To test H1a, b, and c, we sought to determine whether discussion had a larger effect on  
325 the answers than solitary reflection. Figure 1 summarizes the evolution through time of the  
326 different groups with the average response for each problem (Supplementary videos 1 to 6  
327 illustrate this evolution using the spatial layout of the rooms in which the experiment was  
328 carried out; the videos of each group are available in the public repository of the experiment).  
329 Following Claidière et al. (2017), we used mixed models to study the interaction between the  
330 experimental phase (Silence vs. Discussion), and time during the first 10 timesteps (to  
331 maintain the same number of observations in the two phases: 5 in each of the Silence and  
332 Discussion phases). We report the models that combined the problems of each type; however,  
333 we also analyzed each problem independently and found that the results of the combined  
334 models also applied to each problem independently (full reporting of the models can be found  
335 in the Electronic Supplementary Materials). As in our previous study we found that discussion  
336 favored the dissemination of the correct response for the two demonstrative problems ( $\beta =$   
337  $0.38$ ,  $SE = 0.04$ ,  $z = 8.37$ ,  $p < 0.001$ ). For the two factual problems, there was also a  
338 significant interaction between the Silence and Discussion phases, with a reduction in the  
339 distance to the correct response observed only during the Discussion Phase ( $\beta = -2.31$ ,  $SE =$   
340  $0.74$ ,  $df = 6586$ ,  $t = -3.12$ ,  $p = 0.002$ ; see Fig.2).

341



342

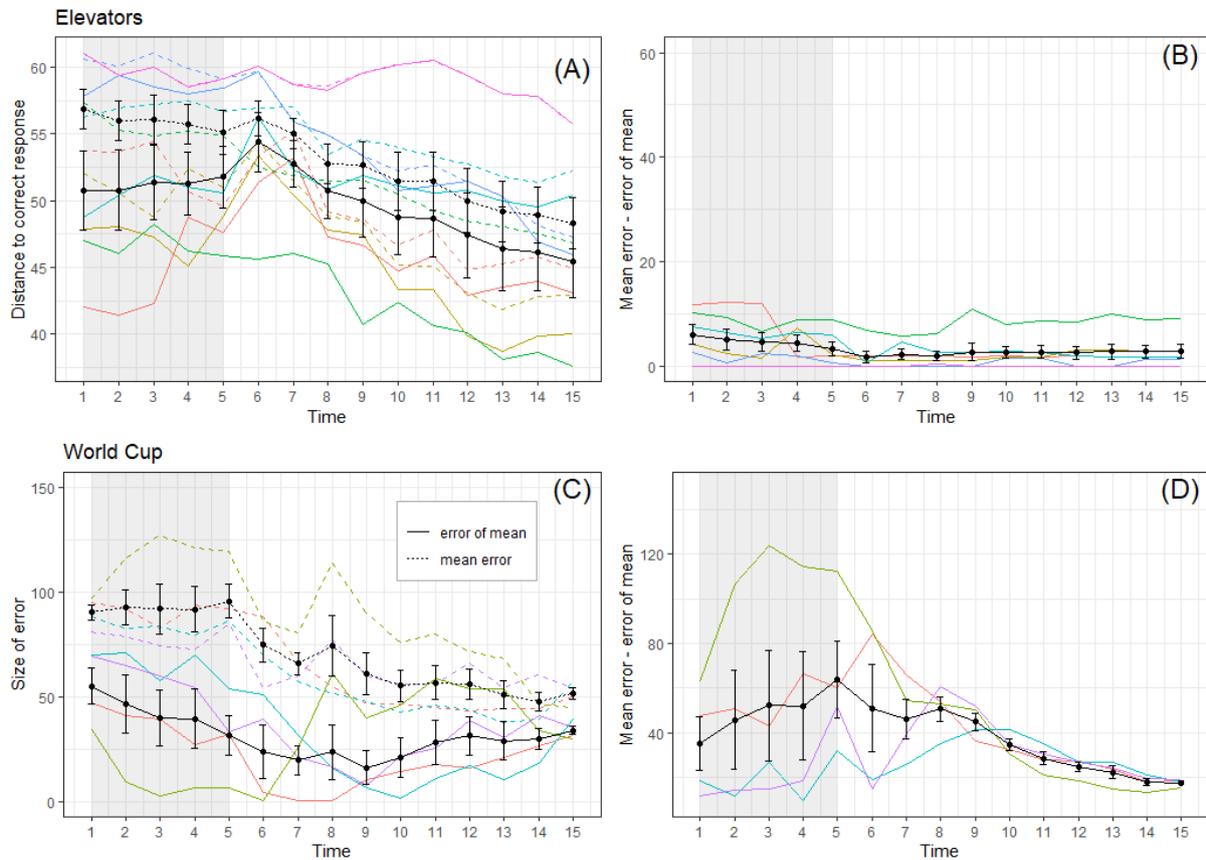
343 Figure 1: Evolution of the group response for each problem through the Silence (shaded area) and  
 344 Discussion phases. Each colored line represents a unique group mean response and the black line  
 345 represents the between group mean (+/- SE). The correct answer to the Elevators problem was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

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

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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



360

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

366

### Discussion

367

H1a, b, and c were confirmed. For both demonstrative and factual problems,

368

discussion improved performance over solitary thinking. The results also clearly supported

369

H2: for demonstrative problems, discussion improved not only individual answers, but also

370

the answers favored by majority voting, which went from two correct answers at the end of

371

beginning of the Discussion Phase, to 13 out of 13 at the end.

372

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

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

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

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

### 409 **Conclusion**

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

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

423 for the problems examined here.

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

432         The positive effects of discussion are also clear for the two factual problems. In the  
433 Elevator problem, all groups correctly increased their average answer through discussion,  
434 moving from a mean error of 55 at the beginning of the discussion to a mean error of 48 at the  
435 end. In the World Cup problem, discussion nearly halved the mean error from 96 to 52. We  
436 also note that asking participants to estimate the number of goals scored in one specific world  
437 cup is a very high bar and it is remarkable that the average number of goals scored in the past  
438 six world cups is 160 goals, a difference of only 19 goals with the grand average reached at  
439 the end of the discussion. Moreover, in our experiments, participants were constrained in  
440 terms of who they could discuss the problems with. Giving people flexibility in network  
441 formation might further increase the advantages of discussion (see, e.g., Almaatouq et al.,  
442 2020). Alternatively, constraining networks to optimize the flow of information has also been  
443 shown to improve accuracy when discussion is not possible, but the same results might extend  
444 to situation in which discussion is possible (Jönsson et al., 2015).

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

## DOES DISCUSSION MAKE CROWDS ANY WISER?

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

452

### 453 **Acknowledgments**

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

464

- 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.

## DOES DISCUSSION MAKE CROWDS ANY WISER?

- 514 Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on  
515 mathematical intellectual tasks. *Journal of Experimental Social Psychology*, 22, 177–189.
- 516 Laughlin, P. R., Zander, M. L., Knievel, E. M., & Tan, T. S. (2003). Groups perform better  
517 than the best individuals on letters-to-numbers problems: Informative equations and effective  
518 reasoning. *Journal of Personality and Social Psychology*, 85, 684–694.
- 519 Liberman, V., Minson, J. A., Bryan, C. J., & Ross, L. (2012). Naïve realism and capturing the  
520 “wisdom of dyads.” *Journal of Experimental Social Psychology*, 48(2), 507–512.
- 521 Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can  
522 undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*,  
523 108(22), 9020–9025.
- 524 Manville, B., & Ober, J. (2003). *A Company of Citizens: What the World’s First Democracy*  
525 *Teaches Leaders About Creating Great Organizations*. Harvard Business Review Press.
- 526 Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D.,  
527 Atanasov, P., Swift, S. A., & others. (2014). Psychological strategies for winning a  
528 geopolitical forecasting tournament. *Psychological Science*, 25(5), 1106–1115.
- 529 Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Trends in*  
530 *Cognitive Sciences*, 20(9), 689–700.
- 531 Mercier, H., Castelain, T., Hamid, N., & Marín Picado, B. (2017). The power of moral  
532 arguments. In J. F. Bonnefon & B. Trémolière (Eds.), *Moral Inferences*. Psychology Press.
- 533 Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative  
534 theory. *Behavioral and Brain Sciences*, 34(2), 57–74.
- 535 Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press.
- 536 Minson, J. A., Liberman, V., & Ross, L. (2011). Two to Tango. *Personality and Social*  
537 *Psychology Bulletin*, 37(10), 1325–1338.
- 538 Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality.  
539 *Thinking and Reasoning*, 4(3), 231–248.
- 540 Moussaïd, M., Herzog, S. M., Kämmer, J. E., & Hertwig, R. (2017). Reach and speed of  
541 judgment propagation in the laboratory. *Proceedings of the National Academy of Sciences*,  
542 114(16), 4117–4122.
- 543 Moussaïd, M., Kämmer, J. E., Analytis, P. P., & Neth, H. (2013). Social influence and the  
544 collective dynamics of opinion formation. *PloS One*, 8(11), e78433.
- 545 Navajas, J., Niella, T., Garbulsky, G., Bahrami, B., & Sigman, M. (2018). Aggregated  
546 knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature*  
547 *Human Behaviour*, 2(2), 126.
- 548 R Core Team. (2020). *R: A language and environment for statistical computing*. R  
549 Foundation for Statistical Computing. <https://www.R-project.org/>
- 550 Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment.  
551 *Organizational Behavior and Human Decision Processes*, 43(1), 1–28.
- 552 Surowiecki, J. (2005). *The Wisdom Of Crowds*. Anchor Books.
- 553 Thorndike, E. L. (1937). Valuations of certain pains, deprivations, and frustrations. *The*  
554 *Pedagogical Seminary and Journal of Genetic Psychology*, 51(2), 227–239.
- 555 Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the  
556 good performance of reasoning groups. *Journal of Experimental Psychology: General*,  
557 143(5), 1958–1971.
- 558 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.
- 559