Experts and laymen grossly underestimate the benefits of argumentation for reasoning

Hugo Mercier

Université de Neuchâtel

Emmanuel Trouche

CNRS &

Université Lyon 1

Hiroshi Yama

Osaka City University

Christophe Heintz,

Central European University

Vittorio Girotto,

University IUAV of Venice

Accepted in Thinking & Reasoning.

Not proofread – please do not quote.

Abstract

Many fields of study have shown that group discussion generally improves reasoning performance for a wide range of tasks. This article shows that most of the population, including specialists, does not expect group discussion to be as beneficial as it is. Six studies asked participants to solve a standard reasoning problem—the Wason selection task—and to estimate the performance of individuals working alone and in groups. We tested samples of U.S., Indian, and Japanese participants, European managers, and psychologists of reasoning. Every sample underestimated the improvement yielded by group discussion. They did so even after they had been explained the correct answer, or after they had had to solve the problem in groups. These mistaken intuitions could prevent individuals from making the best of institutions that rely on group discussion, from collaborative learning and work teams to deliberative assemblies.

Keywords: Reasoning; group problem solving; argumentation; intuitions about argumentation.

Descartes forcefully put forward a view of reasoning as chiefly aimed at improving individual cognition: "the kind of logic which teaches us to direct our reason with a view to discovering the truths of which we are ignorant." By contrast, argumentation—"a dialectic which teaches ways of expounding to others what one already knows"—only "corrupts good sense rather than increasing it" (Descartes, 1985, p. 186). Nineteenth century scholars of crowd psychology attacked even more fiercely institutions relying on deliberation such as juries and parliaments (e.g. Le Bon, 1897), and their views exerted a considerable influence on many 20th century intellectuals (see Barrows, 1981; Moscovici, 1985).

Other, generally less influential thinkers have suggested that reasoning chiefly serves social functions, notably argumentation, and that deliberation is an effective mean to gain better beliefs (Cattaneo, 1864, see Billig, 1996; Landemore, 2012). Many studies have vindicated this minority view by demonstrating that group discussion often improves reasoning performance. This improvement has been observed in a wide range of tasks in the laboratory—deductive problems (Laughlin & Ellis, 1986; Moshman & Geil, 1998; Trouche, Sander, & Mercier, in press), inductive problems (Laughlin, Bonner, & Miner, 2002), numerical estimations (Minson, Liberman, & Ross, 2011; Sniezek & Henry, 1989), and various work related problems (Blinder & Morgan, 2005; Lombardelli, Proudman, & Talbot, 2005; Michaelsen, Watson, & Black, 1989)—as well as in various other contexts—such as work teams (Guzzo & Dickson, 1996), political discussions (Fishkin, 2009; Mercier & Landemore, 2012), scientific discussions (Dunbar, 1995; Mercier & Heintz, forthcoming; Okada & Simon, 1997), and forecasting groups teams (Mellers et al., 2014; Rowe & Wright, 1996). Group discussion yields similar improvements in different cultures (Mercier, 2011a; Mercier, Deguchi, Van der Henst, & Yama, submitted) and throughout development, starting with preschool children (Doise & Mugny, 1984; Mercier, 2011b; Perret-Clermont, 1980; Slavin, 1995; Smith et al., 2009). These results are robust provided some minimal conditions

are met, such as allowing everyone to express their true opinions (Janis, 1982), and providing an heterogeneous opinion pool (Sunstein, 2002).

Although these results are robust, and, in some cases, old (Bos, 1937; Joubert, 1932; Shaw, 1932), they are not mentioned in current reasoning handbooks (e.g. Manktelow, 2012), and, as we have observed in informal discussions, often surprise the general public as well as specialists. Although the view that reasoning works better in deliberative than individual settings has been empirically vindicated, it does not seem to have become dominant, even among specialists. This potential ignorance of the benefits of group reasoning could have dire practical consequences, leading for instance individuals to neglect collaborative learning as an educational method, to underuse teams in organizations, or even to scorn institutions that rely on deliberations such as juries.

In this article we evaluate people's intuitions about the efficacy of group discussion using the most investigated reasoning problem: the *selection task*, in which participants have to evaluate the truth status of a conditional statement (Wason, 1966). In the following studies, after tackling the standard, abstract version of the task, participants were asked to estimate how many people would solve it on their own, and how many would solve it after discussing it in small groups. These estimates could then be compared to the data in the literature which suggest that fewer than 15% of participants working on their own provide the correct answer (Manktelow, 2012), while about 70% do so after discussing in groups of 3 to 5 individuals (see Table 1).

The existing data and the estimates could be compared in two ways. First, one could compare the absolute levels of performance, to determine whether participants can correctly estimate how many individuals get the right answer individually and in groups. Second, one can compare the relative levels of performance—for instance, the ratio of group to individual performance—to determine whether participants can correctly estimate the improvement

yielded by group discussion. Here we are interested in whether participants can anticipate that group reasoning outperforms individual reasoning, not in whether they can correctly estimate absolute levels of performance. Thus, we focus on the second type of comparison, namely, the ratios of group to individual performance.

Source	% individuals correct after solitary reasoning	% groups correct after group discussion	% individuals correct after group discussion	Ratio of group to individual performance
(Moshman & Geil, 1998 comparison 1)	9%	70%	N/A	7.47
(Moshman & Geil, 1998 comparison 2)	21%	80%	79%	3.75
(Maciejovsky & Budescu, 2007)	9%	50%	N/A	5.71
(Mercier et al., submitted)	20%	65%	64%	3.13
Weighted averages	15%	63%	N/A	4.14

Table 1. Comparison of individual and group performance on the selection task. The ratios were computed using the "% individuals correct in groups" when possible.

Study 1

Method

Participants

25 participants (56% women, M_{Age} = 38.28, *SD* = 11.12) were recruited through the Amazon Mechanical Turk website. Their I.P. addresses indicated that they were in the U.S. In Studies 1 to 3, participants were paid the normal rate for this type of task.

Design

The order of the questions 'estimation of individual performance' and 'estimation of

group performance' was counterbalanced.

Procedure

Participants were given the standard, abstract version of the selection task to tackle. Once they had answered, they were asked to estimate individual performance ("Out of 100 people trying to solve this problem on their own, how many people do you think would give the correct answer?") and group performance ("Out of 100 people trying to solve this problem by discussing in small groups, how many people do you think would give the correct answer?"). As a debiasing procedure, participants were then provided with the correct answer to the selection task and its explanation, and they had to estimate individual and group performance again. Finally, they answered standard demographic questions.

Results and discussion

The order of the individual and group estimation questions did not significantly affect the answers in this study or any of the other studies in which it was counterbalanced (Studies 1 to 5). Hence, this manipulation will not be reported in the other studies.

To compare estimated performance with actual performance, we used the four comparisons of individual to group performance that we could locate in the literature (see Table 1), treating each as an individual data point. This N of 4 renders the statistical tests very conservative. In Study 1, individual performance was estimated to be 65% correct (*SD* = 19.76), significantly higher than actual individual performance ($t(13.8) = 9.72$, $p < .001$).[1] Group performance was estimated to be 72% correct (SD = 23.15), not significantly different from actual performance ($t(7.2) = 0.90$, $p = .40$), but significantly higher than estimated individual performance ($t(24) = -3.15$, $p = .004$). The ratios of estimated group to individual performance ($M = 1.12$, $SD = 0.23$) was significantly lower than the observed ratios ($t(3.0) = -3.95$, $p = .029$). Answers following the debiasing procedure will be discussed below, after

[1] The fractional degrees of freedom stem from the use of t-tests on samples with unequal variance.

Study 5. Table 2 presents the main results and Figure 1 presents the ratios of individual to group performance from the present studies.

	Estimated individual performance	Estimated group performance
S1 (U.S.) Before Feedback	65	72
S2 (U.S.) Before Feedback	66	56
S3 (India) Before Feedback	57	62
S4 (Japan) Before Group	59	76
S5 (Managers) Before Feedback	57	71
S1 (U.S.) After Feedback	39	51
S2 (U.S.) After Feedback	52	49
S3 (India) After Feedback	47	46
S4 (Japan) After Feedback	63	75
S5 (Managers) After Feedback	36	57
S6 (Psychologists)	16	38
Global average	49	59

Table 2. Estimated individual and group performance from all studies.

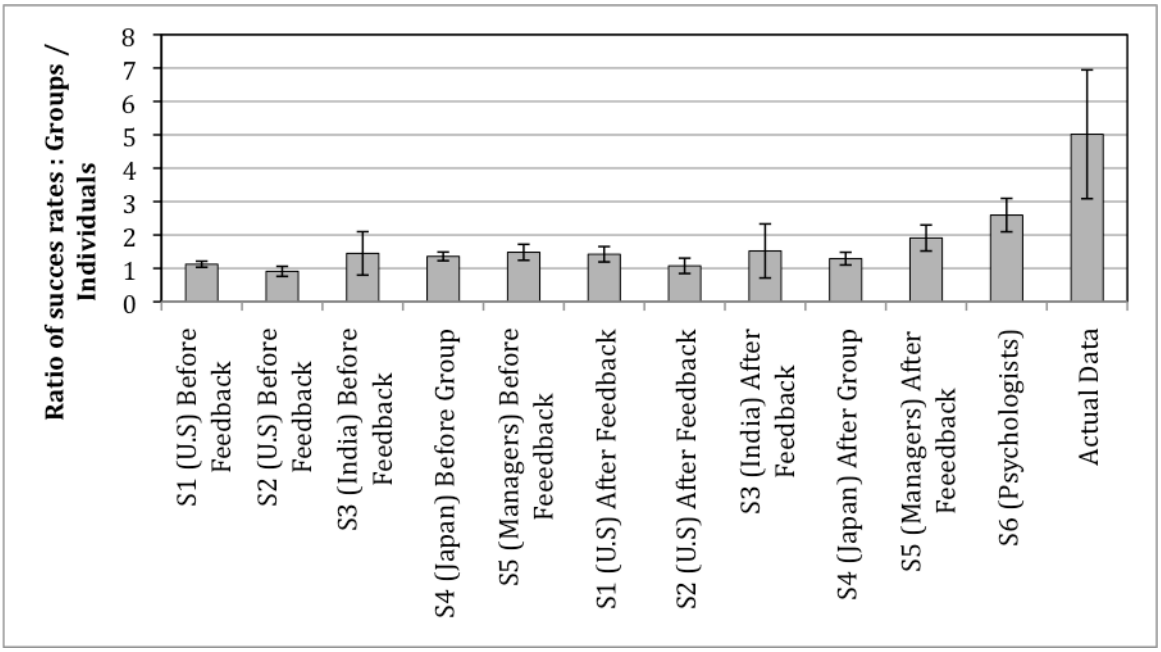

Figure 1. Ratio of group to individual performance: estimates from 6 studies compared with actual data (with 95% confidence intervals).

Previous research has shown that participants fail to appreciate the benefits of aggregating several opinions—averaging opinions in particular—by contrast with choosing one of the opinions (Larrick & Soll, 2006; Soll & Larrick, 2009). In the present case, participants' failure to appreciate the margin by which groups perform better than individuals could stem from difficulties with probabilistic reasoning, namely a failure to compute how many groups would contain at least one member able to find the correct answer on her own. Alternatively, participants could think that even if someone has the correct answer, she will not be able to convince someone with the wrong answer. Given that participants gave a very high estimate of individual performance, the first alternative is unlikely to explain the results, as it would require that there be no mixing of members with the correct and the incorrect answer in any group. Therefore Study 2 focuses on the second explanation, as well as serving as a replication of Study 1.

Study 2

Method

Participants

43 participants (33% women; M_{Age} = 28.0, SD = 4.91) were recruited through the Amazon Mechanical Turk website. They had to be located in the U.S.

Procedure

The procedure was identical to that of Study 1 except that after the estimation questions, participants were asked to directly estimate the effectiveness of argumentation ("Now imagine only two people. One of them has found the correct solution on his or her own, and the other hasn't. The two of them have to agree about an answer. What do you think are the chances that the participant who got the problem right will convince the other? Give an estimate between 0 and 100.").

Results and discussion

The results of Study 1 were replicated. Individual performance was estimated to be 66% correct (SD = 22.18), significantly higher than actual individual performance ($t(11.4) = 10.63$, $p < .001$). Group performance was estimated to be 56% correct (SD = 33.34), not significantly different from actual performance ($t(8.3) = -1.20$, $p = .26$) but significantly lower than estimated individual performance ($t(42) = 2.05$, $p < .05$). As a result, the ratios of estimated group to individual performance ($M = 0.91$, $SD = 0.51$) was significantly lower than the observed ratios ($t(3.0) = -4.15$, $p = .025$).

Participants estimated that someone with the correct answer would convince someone with the wrong answer in 43% ($SD = 25.3$) of the cases. In reality this number is close 100% since the "truth wins" scheme best explains the performance of groups on intellective tasks such as the Wason selection task: As soon as one group member has the correct answer, she nearly always manages to convince the group, even if she is alone and faces a unanimous majority supporting the wrong answer (see for instance Trouche et al., in press). This result confirms that participants grossly underestimate the benefits of argumentation. However, there was no correlation ($r = 0.00$) between the ratios of individual to group performance and the estimation of the efficacy of argumentation in pairs, so that the latter result might not explain the former. Answers following the debiasing procedure will be discussed below.

The underestimation of the benefits of argumentation observed among U.S. participants might reflect the influence of culture specific factors. In particular, Westerners tend to have a more essentialist view of intelligence than Easterners such as than Indian (Rattan, Savani, Naidu, & Dweck, 2012) and Japanese individuals (Heine et al., 2001). An essentialist view of intelligence suggests that intelligence is little affected by learning and other contextual factors (Dweck, 1999) and could therefore explain why U.S. participants do not provide different estimations for individual and group reasoning: they might believe that any individual's chance of providing the correct answer is unaffected by her social setting,

including whether someone else in the group found the correct answer. Accordingly, we replicated Study 1 with participants in India (Study 3) and in Japan (Study 4).

Study 3

Method

Participants

25 participants (36% women; Mage = 33.12, SD = 9.69) were recruited through the Amazon Mechanical Turk website. They had to be located in India.

Procedure

The procedure was identical to that of Study 1 (in English).

Results and discussion

The results of Study 1 were replicated. Individual performance was estimated to be 57% correct (*SD* = 31.17), significantly higher than actual individual performance ($t(23.9) = 5.99$, $p < .001$). Group performance was estimated to be 62% correct (*SD* = 32.07), not significantly different from actual performance ($t(11.6) = -0.44$, $p = .67$) or estimated individual performance ($t(24) = -0.91$, $p = .37$). As a result, the ratios of estimated group to individual performance (*M* = 1.45, *SD* = 1.67) was significantly lower than the observed ratios ($t(3.72) = -3.43$, $p = .030$), and not significantly different from that of U.S. participants ($t(25.25) = 1.36$, $p = .18$). Answers following the debiasing procedure will be discussed below.

Study 4

Method

Participants

35 participants (80% women; *Mage* = 22.8, *SD* = 10.5) took part in the experiment during a class held at Osaka City University. All participants were Japanese. The experiment was not part of the coursework, and students had the option to opt out. The questionnaires

were filled anonymously.

Procedure

The first part of the experiment was identical to that of Study 1, except that the questionnaires were translated into Japanese. Participants had to solve the selection task on their own, and then answer the two estimation questions. As a debiasing procedure, participants were put in small groups and asked to solve the task again. Next, they were asked to answer the two estimation questions again.

Results and discussion

The results of Study 1 were replicated. Individual performance was estimated to be 59% correct (SD = 16.37), significantly higher than actual individual performance ($t(8.1) = 10.19$, $p < .001$). Group performance was estimated to be 76% correct (SD = 17.41), not significantly different from actual performance ($t(4.5) = 1.58$, $p = .18$) but significantly higher than estimated individual performance ($t(34) = -6.77$, $p < .001$). The ratios of estimated group to individual performance (M = 1.36, SD = 0.41) was significantly lower than the observed ratios ($t(3.0) = -3.70$, $p = .034$). However, the ratios were significantly higher than those of U.S. participants (Studies 1 and 2) ($t(73.97) = 4.25$, $p < .001$). Answers following the debiasing procedure will be discussed below.

The results of Studies 1 to 4 suggest that the underestimation of the benefits of argumentation cannot be entirely explained by one critical cultural factor—essentialist thinking about intelligence. They thus suggest that universal mechanisms are at play. However, even if such a cultural factor has no effect on the present results, individual experience with group decision-making might affect the evaluation of individual vs. group reasoning. Managers tend to have extensive experience with team work, and therefore offered a relevant control.

Study 5

Method

Participants

Eighty-six participants took part in the experiment during two classes held at the Central European University (Budapest) as part of an MBA course and an EMBA course. The only data analyzed were from the 46 (35% women; M_{Age} = 35.02, *SD* = 4.51) participants who answered that their current occupation was manager were kept. They had an average of 6.24 years of experience as managers (*SD* = 3.63).

Procedure

Participants answered the questions in a classroom using the online survey of Study 2 with adapted demographic questions (in English).

Results and discussion

The results of Study 2 were replicated. Individual performance was estimated to be 57% correct (SD = 25.29), significantly higher than actual individual performance ($t(13.6) = 8.30$, $p < .001$). Group performance was estimated to be 71% correct (*SD* = 27.25), not significantly different from actual performance ($t(6.10) = 0.80$, $p = .45$), but significantly higher than estimated individual performance ($t(45) = -3.65$, $p < .001$). The ratios of estimated group to individual performance (*M* = 1.48, *SD* = 0.83) were significantly lower than the observed ratios ($t(3.1) = -3.56$, $p = .036$). However, the ratios were higher than those of the non-managers (Studies 1 to 4) ($t(80.88) = -2.10$, $p = .039$) The managers also underestimated the effectiveness of argumentation, judging that someone with the correct answer only had 28 chances out of 100 (*SD* = 20.1) to convince someone with the wrong answer.

Effects of debiasing procedures

The first debiasing procedure, used in Studies 1, 2, 3, and 5, was to explain the correct answer to the selection task. This procedure lowered the estimates of individual (Pre: *M* =

61%, *SD* = 24.8; Post: *M* = 44%, *SD* = 25.8; paired t-test: $t(138) = 7.34$, $p < .001$) and group performance (Pre: *M* = 65%, *SD* = 30.05; Post: *M* = 52%, *SD* = 31.29; paired t-test: $t(138) = 5.47$, $p < .001$). It had little effect on the difference between the estimates of individual to group performance, as this difference remained significant and in the correct direction only for the two groups in which it had the same properties before the debriefing (Study 1, $t(24) = -4.67$, $p < .001$; Study 2, $t(42) = 0.68$, $p = .5$; Study 3, $t(24) = 0.26$, $p = .8$; Study 5 $t(45) = -5.19$, $p < .001$).

The first debiasing procedure had a small but significant positive effect on the ratios of individual to group performance (Pre: *M* = 1.23, *SD* = 0.93; Post: *M* = 1.49, *SD* = 1.30 ; paired t-test: $t(138) = -2.22$, $p = .028$). This effect, however, is entirely driven by the managers (Study 5): Studies 1 to 4 ($t(92) = -1.17$, $p = .244$); Study 5 ($t(45) = -2.37$, $p = .022$). However, the post-debiasing procedure ratios were still significantly lower than the actual ratios ($t(3.08) = -3.55$, $p = .037$), even for the managers ($t(3,2) = -3.09$, $p = .048$).

Explaining to the participants the correct answer had a larger impact on the estimations of the effectiveness of argumentation, possibly because the participants have just been convinced to change their mind in order to adopt the correct answer: Study 2, M_{Pre} = 42%, *SD* = 25.9; M_{Post} = 61%, SD = 21.7, $t(42) = -17.67$, $p < .001$; Study 5, M_{Pre} = 28%, *SD* = 20.3; M_{Post} = 73%, SD = 25.4; $t(45) = -26.76$, $p < .001$. Still, even the post-debiasing estimates were lower than the actual results (close to 100%).

The second debiasing procedure, used in Study 4, was to let participants solve the task in groups. It had no significant effect on the estimates of individual (*M* = 63.29, *SD* = 16.93; paired t-test: $t(34) = -1.28$, $p = .21$), or group performance (*M* = 75.43, *SD* = 15.31; paired t-test: $t(34) = 0.31$, $p = .76$). After the debiasing procedure, the participants still estimated group performance to be higher than individual performance ($t(34) = -3.53$, $p = .001$), but there was no effect of the procedure on the ratios of group to individual performance (*M* =

1.29, *SD* = 0.56; paired t-test: $t(34) = 0.72$, $p = .48$).

The results suggest that the underestimation of the benefits of argumentation is very robust. To further check this conclusion, we tested whether extensive expertise in the psychology of reasoning would allow participants to properly estimate the benefits of argumentation.

Study 6

Method

Participants

Fifty participants were recruited through a professional mailing list (8), personal contacts (27), and at a reasoning workshop (17). We only kept those participants whose self-defined primary field of expertise was psychology of reasoning ($N = 32$) ($M_{Age} = 44.6$, *SD = 13.9*).

Procedure

Participants were told that the object of the study was the Wason selection task, more specifically the standard, abstract version of the task used in Studies 1 to 5. They were then asked to estimate individual and group performance. Participants then had to estimate the effectiveness of argumentation in a simple debating pair, as in Study 2. Finally they answered some demographic questions.

Results and discussion

Participants correctly estimated individual performance ($M = 16\%$, SD = 10.23; $t(4.96) = 0.29$, $p = .78$), and, while they estimated group performance to be higher than individual performance ($t(31) = -7.28$, $p < .001$)., they still underestimated it ($M = 36\%$, SD = 18.28; $t(4.89) = -4.33$, $p = .008$). As a result, they tended to underestimate the ratio of group to individual performance ($M = 2.60$, SD = 1.43; $t(3.40) = -2.38$, $p = .087$). However, they did so less than the other populations, even after they had been given the correct answer

(comparison with the post-feedback ratios of Studies 1, 2, 3, and 5: $t(43.5) = 3.98$, $p < 0.001$).

The psychologists underestimated the effectiveness of argumentation to the same extent that the participants of Studies 2 and 5, answering that someone with the correct answer would convince someone with the wrong answer in only 68% of the cases (SD = 24.15; $M_{\text{Studies 2 and 5}} = 67\%$, $SD = 24.25$; $t(55.01) = -0.17$, $p = .87$).

This result yields two conclusions. First, even experts in the field who are well acquainted with the individual performance on the selection task do not know of the results demonstrating a dramatic improvement after group discussion. Second, these experts do not have the intuition that such a dramatic improvement would take place.

Conclusion

Participants had to solve a standard reasoning problems (except in one study in which it was already known to the participants), and estimate individual and group performance on the same problem. These estimations were compared to the observed performance of individuals and groups in four experiments. All the groups tested underestimated the increase in performance that follows from group discussion (Figure 1). The ratios of group to individual performance were often close to 1, indicating that on average participants thought group discussion would provide no benefits at all over individual reasoning. Indeed, if we exclude the psychologists, we find that before the debiasing procedure over a third of the participants estimated the performance of groups to be the same or lower than that of individuals (65 out of 177 participants). We obtained convergent results when we asked participants to estimate the effectiveness of argumentation more directly by indicating the chances that someone with the correct answer would convince someone with the wrong answer (Studies 2, 5, and 6).

Besides showing that individuals tend to underestimate the benefits of group discussion, our results also suggest that they overestimate individual performance in this type

of task. The participants even kept overestimating individual performance after they had been explained the correct answer—and thus, for most of them, after realizing that they had given the wrong answer. This phenomenon deserves further investigation.

The first moderator to be studied was culture (Studies 1, 2, 3, and 4). We found that the members of cultures that are supposed to have a less essentialist view of intelligence (Indian and Japanese participants) also grossly underestimated the benefits of argumentation. The Japanese participants did so less than the American participants, but this effect could also depend on other differences between the populations (respectively, students vs. MTurkers) and the experimental settings (respectively, in a classroom vs. online).

The second moderator studied was occupation. In Study 5, the participants were managers, people who have experience working in teams and organizing teamwork. They, too, underestimated the benefits of argumentation, although they did so less than other participants. Again, other factors (such as experimental setting) cannot be entirely ruled out as an explanation for this difference.

The third moderator studied was knowledge of the correct answer, which was manipulated as a within-participant variable. The participants for whom this manipulation had the most effect were the managers, and the question for which this manipulation had the strongest effect was the direct estimation of the chances that someone with the correct answer would convince someone with the wrong answer. The latter result can presumably be explained by the fact that the participants had just been convinced to accept the correct answer themselves, and could therefore more easily imagine how correct arguments can modify beliefs. However, the ratios of group to individual performance were less affected, suggesting that participants failed to translate this understanding of the effectiveness of one to one argumentation into more accurate estimations of group performance.

The fourth moderator, also manipulated as a within-participant variable, was solving

the problem in groups (Study 4). Even though performance significantly improved after group discussion (from 20% to 65% correct), the participants did not provide more accurate ratios of group to individual performance after group discussion. The discrepancy with the effects of the previous moderator might stem from the different sources providing the right answer: the experimenter (who is nearly always believed) vs. other group members (who might convince with less certainty).

Finally, the fifth moderator studied was expertise with the task in hand. In Study 6, participants were psychologists of reasoning, whose knowledge of the task was apparent in their correct estimates of individual performance. However, they grossly underestimated group performance, as well as the chances that someone with the correct answer would convince someone with the wrong answer.

These results demonstrate a consistent underestimation of the benefits of group reasoning. It should be stressed, however, that some participants did indicate that groups would perform better than individuals. In particular, both the managers after they had been given the correct answer, and psychologists of reasoning generated ratios of individual to group performance above 1.5. It is therefore possible that experience with the task in hand, coupled with more general expertise about group reasoning, can lead people to correctly estimate that groups perform better than individuals—while still underestimating the size of this effect, as well as, in the case of the psychologists, the efficacy of argumentation in pairs.

A potential concern with the present study is lack of ecological validity, as one might argue that the Wason Selection Task is not representative of everyday reasoning. The Wason Selection Task was chosen thanks to the robustness of its results both in individuals and in groups, making for a sound benchmark. As noted in the introduction, the benefits of group reasoning extend far beyond this and other demonstrative tasks. It would therefore be worthwhile to conduct similar experiments asking participants to estimate individual and

group performance on other reasoning tasks.

The causes of the underestimation of the benefits of group reasoning should be the topic of further study. In any case, these findings suggest that people might be neglecting argumentation as an effective mean of improving a variety of outcomes, from work decisions to school achievement or even political opinions. None of the investigated moderators enabled participants to provide accurate assessments of the benefits of argumentation. Therefore, our results suggest that explicit teaching on this topic might be necessary in order to counteract people's misleading intuitions. Such education could enable individuals to enjoy more of the benefits of argumentation through collaborative learning, work teams, deliberative assemblies, and other institutions that rely on argumentation.

Finally, we would like to stress that these results ought to be of particular interest to specialists of reasoning. These scholars have deployed a substantial amount of ingenuity and energy in trying to improve reasoning performance. Yet they have paid scant attention to group reasoning—arguably the most efficient way of improving reasoning performance. This neglect has been accompanied by a more general neglect of the social uses of reasoning, in particular argumentation. We hope that by pointing out the robustness of the benefits of group reasoning, and by showing that these benefits are far from being intuitive, we might get scholars to pay more attention to the study of group reasoning and argumentation.

Data availability statement

All the data is available at this URL:
https://sites.google.com/site/hugomercier/online%20data%20Experts%20and%20laymen%20grossly%20underestimate%20the%20benefits%20of%20argumentation%20for%20reasoning.xlsx?attredirects=0

Acknowledgements.

We benefitted from an Ambizione grant from the Swiss National Fund (to H.M.) and a PhD grant from the D.G.A. to E.T and a grant from the Italian Ministry of Research (PRIN2010-RP5RNM) to V.G. We thank Mike Oaksford and two anonymous referees for their helpful comments. We also thank the colleagues who, for once, have played the role of participants in Study 6.

References

Barrows, S. (1981). *Distorting mirrors: Visions of the crowd in late nineteenth-century France*. New Haven: Yale University Press.

Billig, M. (1996). *Arguing and Thinking: A Rhetorical Approach to Social Psychology*. Cambridge: Cambridge University Press.

Blinder, A. S., & Morgan, J. (2005). Are two heads better than one? Monetary policy by committee. *Journal of Money, Credit and Banking*, 37, 789-812.

Bos, M. C. (1937). Experimental study of productive collaboration. *Acta Psychologica*, *3*, 315–426.

Cattaneo, C. (1864). Dell'antitesi come metodo di psicologia sociale. *Il Politecnico*, *20*, 262–270.

Descartes, R. (1985). *The Philosophical Writings of Descartes, vol. 1*. Cambridge: Cambridge University Press.

Doise, W., & Mugny, G. (1984). *The Social Development of the Intellect*. Oxford: Pergamon Press.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & Davidson, J.E. (Eds.), *The nature of insight* (pp. 365–395)). Cambridge: MIT Press.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia: Psychology Press.

Fishkin, J. S. (2009). *When the People Speak: Deliberative Democracy and Public Consultation*. Oxford: Oxford University Press.

Guzzo, R. A., & Dickson, M. W. (1996). Teams in organizations: Recent research on performance and effectiveness. *Annual Review of Psychology*, *47*(1), 307–338.

Heine, S. J., Kitayama, S., Lehman, D. R., Takata, T., Ide, E., Leung, C., & Matsumoto, H. (2001). Divergent consequences of success and failure in Japan and North America: An investigation of self-improving motivations and malleable selves. *Journal of Personality and Social Psychology*, *81*(4), 599.

Janis, I. L. (1982). *Groupthink* (2nd Rev.). Boston: Houghton Mifflin.

Joubert, G. J. (1932). *Indiwiduele en Kollektiewe Prestasie, 'n dijdrae tot die experimentele groepsigologie.* Amsterdam: Swets en Zeitlinger.

Landemore, H. (2012). *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton: Princeton University Press.

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

Laughlin, P. R., Bonner, B. L., & Miner, A. G. (2002). Groups perform better than the best individuals on letters-to-numbers problems. *Organizational Behavior and Human Decision Processes*, *88*, 605–620.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology*, *22*, 177–189.

Le Bon, G. (1897). *The crowd: A study of the popular mind*. London: Macmillian.

Lombardelli, C., Proudman, J., & Talbot, J. (2005). Committees versus individuals: An experimental analysis of monetary policy decision-making. *International Journal of Central Banking*, *May*, 181–205.

Maciejovsky, B., & Budescu, D. V. (2007). Collective induction without cooperation? Learning and knowledge transfer in cooperative groups and competitive auctions. *Journal of Personality and Social Psychology*, *92*(5), 854–870.

Manktelow, K. (2012). *Thinking and Reasoning: An Introduction to the Psychology of Reason, Judgment and Decision Making*. Hove: Psychology Press.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., … others. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115.

Mercier, H. (2011a). On the universality of argumentative reasoning. *Journal of Cognition and Culture*, *11*, 85–113.

Mercier, H. (2011b). Reasoning serves argumentation in children. *Cognitive Development*, *26*(3), 177–191.

Mercier, H., Deguchi, M., Van der Henst, J.-B., & Yama, H. (submitted). The benefits of argumentation are cross-culturally robust: The case of Japan.

Mercier, H., & Heintz, C. (forthcoming). Scientists' argumentative reasoning. *Topoi*.

Mercier, H., & Landemore, H. (2012). Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology*, *33*(2), 243–258.

Michaelsen, L. K., Watson, W. E., & Black, R. H. (1989). A realistic test of individual versus group consensus decision making. *Journal of Applied Psychology*, *74*(5), 834–839.

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to Tango. *Personality and Social Psychology Bulletin*, *37*(10), 1325–1338.

Moscovici, S. (1985). *The age of the crowd: A historical treatise on mass psychology*. Cambridge: Cambridge University Press.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking and Reasoning*, *4*(3), 231–248.

Okada, T., & Simon, H. A. (1997). Collaboration discovery in a scientific domain. *Cognitive Science*, *21*(2), 109–146.

Perret-Clermont, A.-N. (1980). *Social Interaction and Cognitive Development in Children*. London: Academic Press.

Rattan, A., Savani, K., Naidu, N. V. R., & Dweck, C. S. (2012). Can everyone become highly intelligent? Cultural differences in and societal consequences of beliefs about the universal potential for intelligence. *Journal of Personality and Social Psychology*, *103*(5), 787.

Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting*, *12*(1), 73–89.

Shaw, M. E. (1932). A comparison of individuals and small groups in the rational solution of complex problems. *The American Journal of Psychology*, *44*(3), 491–504.

Slavin, R. E. (1995). *Cooperative Learning: Theory, Research, and Practice* (Vol. 2nd). London: Allyn and Bacon.

Smith, M. K., Wood, W. B., Adams, W. K., Wieman, C., Knight, J. K., Guild, N., & Su, T. T. (2009). Why peer discussion improves student performance on in-class concept questions. *Science*, *323*(5910), 122.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes*, *43*(1), 1–28.

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 780–805.

Sunstein, C. R. (2002). The law of group polarization. *Journal of Political Philosophy*, *10*(2), 175–195.

Trouche, E., Sander, E., & Mercier, H. (in press). Arguments, more than Confidence, Explain the Good Performance of Reasoning Groups. *Journal of Experimental Psychology: General*.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New Horizons in Psychology: I* (pp. 106–137). Harmandsworth, England: Penguin.