Journal of Experimental Psychology: General
2014, Vol. 143, No. 5, 1958–1971

0096-3445/14/$12.00 http://dx.doi.org/10.1037/a0037099

Arguments, More Than Confidence, Explain the Good Performance of Reasoning Groups

Emmanuel Trouche
Centre National de la Recherche Scientifique, Bron, France

Emmanuel Sander
Université Paris 8

Hugo Mercier
Université de Neuchâtel

In many intellective tasks groups consistently outperform individuals. One factor is that the individual(s) with the best answer is able to convince the other group members using sound argumentation. Another factor is that the most confident group member imposes her answer whether it is right or wrong. In Experiments 1 and 2, individual participants were given arguments against their answer in intellective tasks. Demonstrating sound argumentative competence, many participants changed their minds to adopt the correct answer, even though the arguments had no confidence markers, and barely any participants changed their minds to adopt an incorrect answer. Confidence could not explain who changed their mind, as the least confident participants were as likely to change their minds as the most confident. In Experiments 3 (adults) and 4 (10-year-olds), participants solved intellective tasks individually and then in groups, before solving transfer problems individually. Demonstrating again sound argumentative competence, participants adopted the correct answer when it was present in the group, and many succeeded in transferring this understanding to novel problems. Moreover, the group member with the right answer nearly always managed to convince the group even when she was not the most confident. These results show that argument quality can overcome confidence among the factors influencing the discussion of intellective tasks. Explanations for apparent exceptions are discussed.

Keywords: group reasoning, intellective tasks, argumentation, confidence, problem solving

In a review of the deductive reasoning literature Evans stated, "Logical performance in abstract reasoning tasks is generally quite poor" (Evans, 2002, p. 981). Whether they engage in categorical, conditional or disjunctive reasoning, participants often fail to follow the rules of logic. Performance is also poor on simple but misleading mathematical problems such as those of the Cognitive Reflection Test (Frederick, 2005). The performance of participants solving the same problems in groups stands in sharp contrast to these results. In a dramatic demonstration, Moshman and Geil (1998) found that 70% to 80% of participants were able to solve the standard version of the Wason selection task when asked to discuss it in small groups. Individually, the same participants (or participants from the same population) performed in line with previous results, with a rate of good answers of 10% to 20% (see also Maciejovsky & Budescu, 2007; Mercier, Deguchi, Van der Henst, & Yama, 2014).

Beyond the Wason selection task, experiments have shown that for many intellective tasks—"problems or decisions for which there exists a demonstrably correct answer within a verbal or mathematical conceptual system" (Laughlin & Ellis, 1986, p. 177)—group discussion significantly improves performance. This is true of mathematical problems for adults (Laughlin & Ellis, 1986) and of various mathematical or logical problems for children (Doise & Mugny, 1984; Miller & Brownell, 1975; Perret-Clermont, 1980). The scheme that best describes group performance is "truth wins": If a group member has understood the problem, her answer ends up being adopted by the group, even if she is the only one defending it.

A common interpretation of these findings is that group members exchange arguments and the arguments exposed by the participants who have best understood the problem prove to be the most convincing (Moshman & Geil, 1998; Nussbaum, 2008). This interpretation is supported by the analysis of transcripts, which suggests that participants change their minds when they understand the arguments supporting the correct answer (e.g., Trognon, 1993), and by transfer effects. Such transfer effects are the expected outcomes of an abstract encoding of the initial problem

This article was published Online First June 9, 2014.

Emmanuel Trouche, Laboratoire Language, Cerveau et Cognition, Centre National de la Recherche Scientifique, Bron, France; Emmanuel Sander, Laboratoire Paragraphe, Université Paris 8; Hugo Mercier, Centre de Sciences Cognitives, Université de Neuchâtel.

This work was funded by an Ambizione fellowship from the Swiss National Fund to Hugo Mercier, by a grant of the Direction Générale de l'Armement to Emmanuel Trouche, and by a university grant of the Paragraph Lab to Emmanuel Sander. We thank Vittorio Girotto, Jean-Baptiste Van der Henst, and David Moshman for their very useful comments. We also thank the students, teachers, and administrative personnel, in particular the academic advisor Emmanuel Paul, who collaborated with us on Experiment 4.

Correspondence concerning this article should be addressed to Hugo Mercier, Centre de Sciences Cognitives, Université de Neuchâtel, Espace Louis Agassiz 1, 2000 Neuchâtel, Switzerland. E-mail: hugo.mercier@unine.ch

when properly understood, that would apply as well to the transfer problems (Gamo, Sander, & Richard, 2010; Sander & Richard, 1997). Several experiments have demonstrated that the gains in performance during group discussion transfers to other problems that have the same structure but a different appearance (for review, see Laughlin, 2011).

This interpretation, however, has not gone unchallenged. In particular, participants' relative confidence in their answers seems likely to influence the outcome of group discussion. That confidence plays such a role is known for problems closer to the judgmental side of the spectrum (the opposite of intellective problems). When taking others' opinions into account on various numerical estimation tasks, people are known to rely on a "confidence heuristic" (Price & Stone, 2004; see also, e.g., Van Swol & Sniezek, 2005; Yaniv, 1997) such that more confident opinions tend to weigh more. This confidence heuristic also plays an important role in the discussion of judgmental problems. It has been shown, for instance, that in the case of perceptual or even general knowledge questions, the outcome of group discussion can be emulated by aggregating the individual judgments of the group members weighed by their confidence (Koriat, 2012).

It is less obvious that confidence should also play a role in the discussion of intellective problems. Substantial evidence, however, suggests that it does. For instance, when Zarnoth and Sniezek (1997) had dyads solve a series of problems on the judgmental-intellective spectrum, they even found that "the extent to which group members' confidence predicted their influence was also greatest on intellective rather than judgmental tasks" (p. 345). It is important here to distinguish two ways in which confidence influences the outcome of group discussion for intellective tasks. The first is through a correlation between confidence and accuracy. To the extent that this correlation is positive, the effects of confidence and accuracy are confounded. The second is through sheer confidence: Even when they are wrong, confident members could be more likely to influence their peers. Zarnoth and Sniezek (1997) found evidence for both factors. For instance, in easy mathematical problems—canonical intellective problems—out of 24 dyads in which the most confident member was correct, 24 adopted her answer. However, out of 12 dyads in which the most confident member was incorrect, seven also ended up adopting this incorrect answer. More generally, statistical analyses suggest that confidence played a role beyond its correlation with accuracy in these intellective tasks, either in dyads or in larger groups.

Using similar problems, Aramovich and Larson (2013) obtained convergent results. Their analyses of the outcome of group discussions revealed that in the individual answers, "correctness was significantly correlated with confidence," but that, during the discussion, "confidence was clearly the more important variable, affecting participant's group problem-solving preferences regardless of the correctness of their answers" (p. 42). They also discovered one way in which confidence influences group outcomes: by precluding participants who are the only ones to have the good answer but who are not very confident to express their views. Previous studies had found similar overall results using other problems (Johnson & Torcivia, 1967, which is discussed at greater length in the conclusion) or variables that tend to correlate with confidence: expertise (Littlepage, Schmidt, Whisler, & Frost, 1995) and dominance (Anderson & Kilduff, 2009).

Finally, Levin and Druyan (1993) obtained similar results with children. They observed that the children who were able to solve a given Piagetian problem were often able to convince their peers of their answer but that the children with the correct answers were also initially more confident, making of confidence a potential alternative explanation for the outcome of group discussions. Another experiment suggested that confidence was in fact the main driver of group discussion: For other intellective problems, the more confident children were also able to convince their peers, even though they had the wrong answer (see also Tudge, 1989).

This overview of the literature makes it clear that both argument quality and confidence play a role in the discussion of intellective problems. The goal of this article is to provide novel evidence for the respective roles of these two factors. More specifically, it seeks to find a limiting case: problems for which arguments are most likely to play the key role. This provides a strong test of the hypothesis that confidence can account in large part for the discussion of intellective problems. If confidence is found to play a major role even in these cases, it means it is likely to play an at least equally important role in the discussion of less clear cut intellective problems. No such inference could be drawn if arguments were found to play the major role. However, if this were the case, we would have an important demonstration of argumentative competence (by opposition with performance; see Chomsky, 1965), namely, that people have the ability to evaluate arguments properly for their content and mostly disregard other cues such as confidence.

While argument quality and confidence are likely to both play a role in the discussion of intellective problems, it is possible to make a series of predictions derived, respectively, from the *Argument Explanation*—the exchange of arguments is the main factor explaining the performance of groups in intellective tasks—and the *Confidence Explanation*—participants' confidence is the main factor explaining the performance of groups in intellective tasks.

The *Argument Explanation* makes the following predictions:

A1. *Good arguments change people's mind*. They do so even if the arguments are not accompanied by markers of confidence or epistemic authority.

A2. When participants understand the arguments supporting the good answer to a problem, they might be able to recreate these arguments when presented with an analogous problem. *People are sometimes able to transfer the understanding they have gained during group discussion to other tasks*. Crucially, this prediction is unique to the Argument Explanation and could not be derived from the Confidence Explanation.

A3. Since, in intellective tasks, someone with the correct answer can typically muster stronger arguments than someone with the wrong answer, *being right is a better predictor of one's ability to convince one's peers than being confident*.

The *Confidence Explanation* makes the following predictions:

C1. Since we know that participants who have the good answer in intellective tasks tend to convince their peers

("truth wins"), *participants with the good answer are more confident*. While this prediction is crucial for the Confidence Explanation, it is entirely compatible with the Argument Explanation.

C2. *The most confident participants are the least likely to change their minds*. Otherwise the answer of the high confidence member would not be more likely to be adopted by the group.

C3. When being right and being more confident do not correlate perfectly, *being confident is a better predictor of one's ability to convince one's peers than being right*.

In order to properly test these predictions, it is important to take stock of a possible confound present in some of the experiments reviewed above. As suggested by Aramovich and Larson (2013), in a multiple choice answer format, the correlation between having the correct answer and having understood the problem needs not be perfect. Indeed, if some participants answer at random, they will sometimes provide the correct solution. These participants would likely have low confidence and would therefore constitute seemingly crucial cases to tease out the roles of confidence and accuracy (and the associated argument quality). However, the Argument Explanation does not predict that these people should be able to convince their peers. Indeed, these participants would violate one of the conditions for demonstrability—"the correct member must have sufficient ability, motivation, and time to demonstrate the correct solution to the incorrect members" (Laughlin & Ellis, 1986, p. 180)—since one can hardly expect someone who provided an answer by chance, or for completely wrong reasons, to demonstrate the correct solution.

Accordingly, we have relied only on problems for which we could reasonably insure a near perfect correlation between providing the right answer and having properly understood the problem. In Experiment 1, we used problems that had a free-response format (as suggested by Aramovich & Larson, 2013) so that very few, if any, participants would provide the correct answer without having understood the problem. In the following experiments, we relied on a coding of the participants' justifications to ensure that those who had provided the right answer had not done so by chance, or based on a severe misunderstanding of the problem, but because they had grasped its logic.

The following four experiments test the six hypotheses following from the Argument and the Confidence Explanations, either by providing participants with a single argument aimed at changing their minds about an intellective problem (Experiments 1 and 2) or by making participants discuss an intellective problem in groups (Experiments 3 and 4).

Experiment 1: Effect of a Single Argument on Intellective Tasks

Previous results have already provided support for A1 (*good arguments change people's mind*). For instance, Stanovich and West (1999) gave participants a series of intellective tasks such as the sunk cost problem or Newcomb's problem. After the problems, participants were provided with arguments devised by the experimenters, one for the normative answer and one for the nonnormative answer. After being exposed to these arguments, on most problems more participants shifted from the nonnormative to the normative answer than the other way around (see also Slovic & Tversky, 1974).

If one wanted to extrapolate from this individual success to the good performance of groups, however, one would face the problem that the normative arguments used in these experiments were not formulated by participants, as they would be in a group discussion, but by the authors, experts in the relevant domain. It is therefore possible that participants could be swayed by these sound, well-formulated arguments for the normative answer but would remain unaffected by arguments for the same answer provided by other participants. By contrast, the experimenters might have had more difficulties creating equally sound arguments for an answer they knew to be incorrect. To circumvent this limitation, in Experiment 1 we provided participants with the arguments previously formulated by other participants who had to solve the same intellective tasks.

Method

Participants and design. The participants were 42 students recruited in a university in Paris, France (19 women, $M_{Age} = 25.1$ years, $SD = 7.0$). Each participant solved three intellective problems, in a counterbalanced order.

Materials and procedure. The problems used were the three problems of the Cognitive Reflection Test (Frederick, 2005). One of these problems is known as the "Bat and Ball": "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents." The correct answer is "5¢," and the wrong intuitive answer is "10¢." In a preliminary study, a group of participants was asked to solve the problems and to justify their answers verbally. Their justifications were recorded and, for each problem, four of these justifications—two for the correct answer and two for the intuitive but wrong answer—were selected, lightly edited and printed. The arguments selected had to be correct arguments in the case of the correct answer, and all four arguments were approximately matched in length, number of clauses, and number of connectives. For instance, one of the arguments for the correct answer to the Bat and Ball read (translated from French): "If we have a ball at 5¢ and a bat at $1.05, we get 1 dollar more for the ball, and the total is $1.10." And for the intuitive but wrong answer: "10¢, since the bat costs $1, and since the sum is $1.10, the ball has to cost 10¢."

The participants of the experiment were asked to solve one problem and to provide a verbal justification. They were then presented with one of the four arguments previously gathered, selected at random, and given the opportunity to change their minds. They were told "Last week, we conducted the same experiment with other participants. They had to do exactly the same task, so that included recording their justifications for their answers. One of the participants answered *X*. Here is the justification that this participant gave," and shown a sheet of paper with the argument printed on it. The procedure was repeated for the other two problems.

Results and Discussion

Results. There was only one occurrence of a participant changing her mind after being presented with an argument sup-

porting her answer, so we focus on the cases in which the participant was presented with an argument against her answer. Table 1 summarizes these results. It presents the total number of occurrences as well as the number of participants for whom the relevant outcome happened at least once. Fisher exact tests (all tests are two-tailed throughout) confirmed that participants were significantly more likely to change their minds when they had given the wrong answer and been provided with arguments for the correct answer than vice versa (comparing the total number of occurrences, $p < .001$; comparing the participants to whom each outcome happened at least once, $p < .001$). As a matter of fact, no participant who had the correct answer and was presented with an argument for the wrong answer changed her mind.

Discussion. In 45% of the cases, participants who had given the wrong answer and had been presented with an argument for the correct answer changed their minds to adopt the correct answer. This happened despite the fact that the arguments source—another participant—had no epistemic authority and that there were no markers of confidence, either explicit (no argument contained markers, such as "it's obvious" or, conversely, "I guess"), or implicit (since the arguments had been transcribed and then printed, there was no tone, facial expression, etc.). These results are in line with previous observations showing that for many intellective problems arguments for the normative answer are more convincing that arguments for a nonnormative answer (Stanovich & West, 1999). Experiment 1 extends these results by showing that this is true even if the arguments are formulated not by experts but by other participants. A1 is supported: Even without markers of confidence or epistemic authority, good arguments can change people's minds on intellective tasks.

Experiment 2a: Effect of Confidence and of a Single Argument on Intellective Tasks

Experiment 1 could only test Hypothesis A1. Experiment 2a relied on a similar design but asked participants to provide measures of confidence, enabling tests of C1 (*participants with the good answer are more confident*) and C2 (*the most confident participants are the least likely to change their minds*). However, typical measures of confidence—such as asking people to indicate how confident they are that their answer is correct—might not adequately capture the feelings of confidence most relevant in the case of intellective problems. In deceptive intellective tasks (such as those of the Cognitive Reflection Test for instance), people might be very confident in the intuitive but wrong answer. Even if those who give the wrong answer are less confident than those providing the correct answer, they still have very high levels of confidence—for instance, while De Neys et al. (2013) reported 95% average confidence for participants solving the Bat and Ball, those who provided the intuitive but wrong answer were also highly confident (above 80%). However, the participants giving the wrong answer might be less confident in the reason for this answer. Confidence in one's reason, rather than confidence in one's answer, might be the relevant feeling of confidence to predict the outcome of group discussions on intellective tasks. Accordingly, after the participants had provided their answer and estimated their confidence in the answer, they were asked to give a reason for this answer and to estimate their confidence in this reason.

Method

Participants and design. Two hundred fifteen participants were recruited online through Amazon Mechanical Turk (99 women; $M_{Age} = 35.4$ years, $SD = 12.8$). Each solved one intellective problem and was presented with an argument against her answer.

Materials and procedure. The problem used was a disjunctive reasoning task borrowed from Levesque (1986). It read:

> Paul is looking at Linda and Linda is looking at Patrick. Paul is married but Patrick is not. Is a person who is married looking at a person who is not married? Yes/No/Cannot be determined.

The correct answer is "Yes" (if Linda is married, she is looking at Patrick, who is not married; if Linda is not married, Paul, who is married, is looking at her). However, previous experiments (Toplak & Stanovich, 2002) indicate that "Cannot be determined" is the modal answer among typical participants.

Participants were asked to solve the problem and to evaluate their confidence in the correctness of their answer by choosing from 9 indicators of confidence. To avoid ceiling effects due to overconfidence, the scale was skewed to include several answers denoting strong confidence (inspired by Kuhn & Lao, 1996) so that it ranged from "Not confident at all" to "As confident as in the things I'm most confident about" (see Appendix A).

Participants then had to justify their answer. Pilot studies showed that if participants were simply asked to evaluate the confidence in the correctness of their reason on a scale similar to the scale used for the confidence in the correctness of the answer, many participants simply provided the same answer. To lower the artificial correlation between the two measures of confidence, participants were asked to judge the quality of their reason (rather than their confidence) by choosing one of seven (rather than nine) indicators, again skewed to avoid ceiling effects so that they ranged from "Very poor reason" to "A demonstrative reason that perfectly supports its conclusion."

Pilot studies also showed that some participants (around 5%) changed their minds in the process of providing a justification. Accordingly, as they justified their answer, participants were given the possibility to provide a new answer as well as a new measure

Table 1
Number of Occurrences of Participants Changing Their Mind When Confronted With Contrary Arguments in Experiment 1

Variable	Change of mind	No change of mind
Wrong initial answer, argument for the correct answer	18 (16) (adopt correct answer)	22 (17) (stick with wrong answer)
Correct initial answer, argument for the wrong answer	0 (0) (adopt wrong answer)	35 (23) (stick with correct answer)

Note. In parentheses is indicated the number of participants to whom each outcome happened at least once.

of confidence in the correctness of their answer. These answers were used as the baseline of initial answers. Participants were then given an argument against this initial answer, described as having been gathered from another participant in a previous phase of the experiment. Participants who had given the right answer were provided with an argument for the intuitive but wrong answer; participants who had given an incorrect answer were provided with an argument for the right answer. The arguments used were lightly edited versions of arguments given by participants in a pilot study. Participants could give a final answer after the argument. Finally, participants were asked to justify their final answer in order to make sure that if they changed their minds, they did it because they had understood the argument.

Results and Discussion

Results. As mentioned above, what is relevant for the present purposes is not simply having provided the "Yes" answer, but having understood the task. Accordingly, answers were coded as correct only if participants answered "Yes" and gave a correct justification (nine participants provided other justifications). The coding was straightforward as the justifications were either clearly correct (e.g., "if Linda is married she is looking at Patrick, if not Paul is looking at her") or clearly incorrect (e.g., "Because Patrick isn't married and Linda was looking at him"). This means that the wrong justification did not reflect a good understanding of the logic of the task accompanied by an inability to explain it, but a severe confusion, often about the premises. In order to keep the analyses of the participants who had provided the correct answer and the wrong answer as symmetrical as possible, we also eliminated those participants who either answered "No" ($N = 12$) or answered "Cannot be determined" for nonstandard reasons ($N = 9$). Again, the coding was straightforward as the justifications were either clearly standard (e.g., "It is not stated whether Linda is married or not. ") or clearly nonstandard (e.g., "Because we don't know if Patrick is looking at Linda"). As a result, all the analyses below bear on 185 participants, 19 having properly understood the problem, and 166 having provided the standard, intuitive wrong answer.

None of the 19 participants who had given the correct answer changed their minds after being presented with an argument for the wrong answer. By contrast, among the 166 participants who had given the wrong answer. 65 (39%) changed their minds when exposed to an argument for the correct answer. Crucially, 62 of these participants gave an appropriate final argument, so that there were significantly more participants who gave the correct justification among those who had given the wrong answer and then been exposed to a good argument (62 out of 166) than among people being simply confronted with the problem (19 out of 215; Fisher exact test, $p < .001$).

All confidence measures were normalized to 0–100%. Participants who had given the correct answer rated the confidence in their answer on average at 84.9% ($Mdn = 87.5$, $SD = 15.9$) and the confidence in reason at 82.5% ($Mdn = 83.3$, $SD = 21.9$). Those who had given the wrong answer rated the confidence in their answer at 75.0% ($Mdn = 75.0$, $SD = 25.0$) and the confidence in their reason at 75.1% ($Mdn = 83.3$, $SD = 23.6$). Wilcoxon rank sum tests indicated that participants with a wrong answer were not significantly less confident in their answer than those with the right answer (for the confidence in answer, $W = 1286$, $p = .175$; for the confidence in reason, $W = 1280$, $p = .165$), a result that might be due to the small sample of participants providing the correct answer.

To determine whether confidence in answer can predict who changes their minds, the population of participants who had provided the intuitive but wrong answer ("Cannot be determined") was split at the median based on initial confidence in answer ratings. Participants from the lower half of the distribution were not significantly more likely to change their minds (36%) than those from the upper half (42%; Fisher exact test $p = .42$). Regarding confidence in reasons, participants from the lower half of the distribution were significantly *less* likely to change their minds (30%) than those from the upper half (46%; $p = .037$).

Discussion. Experiment 2a confirms the results of Experiment 1 regarding A1: Even without any confidence markers, good arguments can change people's mind in intellective tasks. By contrast, the predictions of the Confidence Explanation were weakly supported (C1) or not supported (C2). Regarding C1, there was a trend toward higher confidence for participants having provided the right answer. This weak difference is in line with previous results that found little or no correlation between confidence and accuracy on reasoning tasks (e.g., Shynkaruk & Thompson, 2006). Experiment 2a offers a strong test of C2, with a large sample of participants who provided the same wrong answer and were confronted with the same argument for the correct answer. There was no relation between confidence in answer and likelihood of changing one's mind, and the opposite relation to that predicted by C2 for confidence in reason.

Experiment 2b: Three Controls for Experiment 2a

In order to strengthen the interpretation of Experiment 2a offered above, we conducted three control experiments. In order to more closely replicate previous results (Stanovich & West, 1999), in Control 1 we exposed each participant to two arguments, one for the correct answer and one for the incorrect answer.[1]

It could be suggested that, in Experiment 2a, what makes participants change their minds is not the logic of the argument but superficial features of its phrasing. Accordingly, we devised a control problem which had the same superficial content (married and unmarried people looking at each other), but a different logic, so that now the correct answer was "Cannot be determined," while "Yes" was incorrect (Control 2). Participants were given the same arguments as in Experiment 2a (presented as in Control 1). If it is the logic of the good argument that makes people change their minds in Experiment 2a, then they should not change their minds in this case, since the logic is not valid anymore. By contrast, if its superficial features make the good argument appealing, then it should still be equally appealing.

In Experiment 2a, the arguments presented to the participants had been picked by the experimenters. Although this is arguably better than if they had been written by the experimenter, this still

[1] We also tried to implement a transfer task in Control 1. However, all but one of the participants who had initially answered the problem correctly failed to solve the transfer task, suggesting that it was poorly calibrated. Experiments 3 and 4 offer more persuasive evidence about transfer tasks.

leaves open the possibility that the experimenters unwittingly picked arguments that were more convincing (for the arguments supporting the correct answer) or less convincing (for the arguments supporting the incorrect answer) than the average. Accordingly, in Control 3, we randomly selected from the arguments generated in Experiment 2a—five arguments for the correct answer (among those who had understood the problem) and five arguments for the incorrect answer (among those who gave the standard justification)—and replicated Experiment 2a with these arguments.

Method

Participants. Two hundred four participants were recruited online through Amazon Mechanical Turk (Control 1 = 60, Control 2 = 30, Control 3 = 114; total of 84 women; M_{Age} = 33.1 years, SD = 11.0).

Materials and procedure. Control 1 is identical to Experiment 2a, except that instead of presenting participants with a single argument against their answer, we presented all the participants with the same arguments: the argument for the correct answer and the argument for the wrong intuitive answer used in Experiment 2a, order counterbalanced.

Control 2 is identical to Control 1, except that the problem used inverted which answer is correct and which is incorrect:

> Paul is looking at Linda and Linda is looking at Patrick. Paul is married but Patrick is not married. Is a person who is married looking at a person who is also married?

Control 3 is identical to Experiment 2a, except that instead of one handpicked argument for the correct and incorrect answers, we used five randomly picked arguments for each answer. Each participant only saw one of the arguments going against her initial answer.

Results and Discussion

Results.

Control 1. Three participants answered "Yes" and two answered "Cannot be determined" for wrong or nonstandard reasons; they were excluded from further analyses. Five participants provided the right answer and the right justification. None of these five participants changed their minds after being presented with the arguments. By contrast, among the 50 participants who had given a wrong answer, 21 (42%) changed their minds when exposed to an argument for the correct answer, leading to significant improvement in performance after the arguments (Fisher exact test, $p < .001$).

Participants who had given the correct answer rated the confidence in their answer on average at 87.5% (Mdn = 87.5, SD = 12.5). Those who had given the wrong answer rated the confidence in their answer at 68% (Mdn = 75.0, SD = 21.0), a significant difference (W = 57, p = .042).

Splitting the 50 participants who had provided the intuitive but wrong answer at the median based on initial confidence showed that participants from the lower half of the distribution were not significantly more likely to change their minds (35%) than those from the upper half (48%; Fisher exact test, p = .40). The same analyses were performed using confidence in reason and lead to the same result.[2] Participants from the lower half of the distribution were not significantly more likely to change their minds (35%) than those from the upper half (52%; Fisher exact test, p = .36).

Control 2. Out of the 26 participants who selected the correct answer ("Cannot be determined"), one (4%) changed her mind when confronted with the arguments, significantly fewer than in Control 1 (Fisher exact test, $p < .001$).

Control 3. Four participants answered "Yes," seven answered "Cannot be determined" or "No" for wrong or nonstandard reasons; they were excluded from further analyses. Twenty participants selected the correct answer with a good justification; none changed her mind. Out of the 83 participants providing the wrong answer with the standard justification, 41(49%) changed their minds for the good answer, and 38 were able to provide a good final justification, more than in Experiment 2a, even if nonsignificantly (Fisher exact test, p = .27). For the five arguments, the percentages of participants changing their minds were 50%, 47%, 63%, 33%, and 53%, respectively. The largest difference was between Argument 3 and Argument 4, still a nonsignificant difference (Fisher exact test, p = .17).

Discussion. The results of all three control studies confirm our interpretation of Experiment 2a. Control 1 replicates all the findings of Experiment 2a using a slightly modified paradigm in which all the participants were confronted with the same arguments. In Control 2, only one participant changed her mind when the argument for the correct answer used in Experiment 2a and Control 1 was used to argue for an incorrect answer. This demonstrates that superficial features of the argument cannot explain why people changed their minds in Experiment 2a or Control 1. Finally, Control 3 shows that the arguments used in Experiment 2a were not selected to be especially good (argument for the correct answer) or poor (argument for the wrong answer). A random sample of five other arguments was, on average, equally effective at convincing people to adopt the correct answer. Indeed, the argument used in Experiment 2a was the second least effective of the six arguments. Although the variations in argument strength are not very reliable due to the small samples, nothing in these results suggests that the arguments selected in Experiment 2a were particular in any way.

Experiment 3: Effects of Confidence and Arguments in Group Discussion Among Adults

A limitation of Experiments 1 and 2 is that they provide relatively impoverished conditions for both arguments and confidence to influence participants' answers. In a discussion, people can reformulate their arguments and offer counterarguments. By contrast, in Experiments 1 and 2 participants were provided with a single argument for the good answer. This could explain the fact that half of the participants did not change their minds, whereas in group discussions of intellective tasks, participants with the wrong

[2] Some participants who had changed their minds in the process of justification interpreted the question about the confidence in reason as bearing on their original answer instead of their new answer. As a result, the answers to the confidence in reason question of the nine participants who had changed their minds in the process of justification were removed from the analysis.

answer exposed to the right answer change their minds much more often (as entailed by the strength of the "truth wins" scheme). However, a variety of verbal and nonverbal markers of confidence can also be expressed in a face-to-face discussion. Participants with the right answer might simply rely on these markers to impose their views, rather than on the back and forth of argumentation. As a result, confidence could take precedence over argument quality in face-to-face discussions.

In Experiment 3, participants faced the same intellective task individually and then in groups before individually solving a transfer task. When first confronting the task, participants had to evaluate the confidence in their answer, provide a reason for their answer, and evaluate the confidence in this reason. Accordingly, Experiment 3 provided a test of all six predictions made by the Argument Explanation and the Confidence Explanation.

Method

Participants and design. The participants were 98 students in Communication and Information Sciences from the University of Neuchâtel, Switzerland (74 women; M_{Age} = 21.1 years, SD = 2.2). The experiment was conducted during class. Participants solved a disjunctive reasoning task individually and then in groups. Finally, they individually solved a transfer task.

Materials and procedure. The task was the same disjunctive reasoning problem as in Experiment 2, translated in French. First, people were asked to solve the problem, to provide a measure of confidence (eight choices, ranging from "Not at all sure" to "So sure that I could never change my mind"), to provide a justification for their answer, and to evaluate this reason (six choices, ranging from "I think these are not good reasons at all" to "I think these are the best possible reasons"). Then 25 groups of three to five participants were formed and asked to solve the same problem until they had reached a consensus. Finally, participants had to solve two transfer problems. The first transfer problem was analogous to the initial problem, while the second was modified so that the correct answer was "Cannot be determined" (see Appendix B) Accordingly, if a participant provided the correct answer to both transfer problems, she was likely to have genuinely understood how to solve these disjunctive reasoning problems, and she was not relying on a simple heuristic of answering "Yes."

Results and Discussion

Results. As in Experiment 2, an answer was coded as correct only when an appropriate justification was provided. Two participants chose the answer "yes" without any appropriate justification. As in Experiment 2, they have been excluded from all the analyses, along with the participants who were in their groups.[3] Nineteen (out of 87, or 22%) participants provided the correct answer individually, and 55 (63% or 14 groups out of 22) provided the correct answer after group discussion. This difference is significant whether we treat each participant, McNemar's chi-squared: $\chi^2(1, N = 87) = 36$, $p < .001$, or each group (Fisher exact test, $p < .001$) as a data point for the postdiscussion answer (see Figure 1).

The "truth wins" scheme described all the groups to which it could apply (i.e., the 11 groups in which at least one member had initially answered correctly). In five groups a single participant with the correct answer convinced her peers, even though they all agreed on the same answer. Three groups also found the correct answer even though all of their members had initially answered "Cannot be determined."

Initial confidence measures (in answer and in reason) were marginally related with correctness. The average confidence in answer for the correct answers was 73.7% (Mdn = 85.7, SD = 23.0) and 63.4% for the wrong answers (Mdn = 57.1, SD = 25.7). For the confidence in reason, the respective averages were 69.5% (correct answer, Mdn = 80.0, SD = 22.5) and 63.8% (wrong answer, Mdn = 60.0, SD = 21.9). Wilcoxon rank sum tests didn't show significant differences, for the confidence in answer (W = 490, p = .10) and for the confidence in reason (W = 552, p = .32). As in Experiment 2a, the lack of statistical significance might be due to the small sample of participants providing the correct answer.

Even though the group members who had the correct answer before group discussion tended to be more confident than their peers, the correlation was far from perfect. In 11 groups one or more members had found the correct answer individually. In all of these groups, they convinced their peers. The effects of confidence can be measured either by averaging over group members with the same answer, or by looking at the individual group member with the highest confidence. In two out of these 11 groups, the member(s) who had the correct answer were not, on average, more confident than their peers. In four out of these 11 groups, the (or one of the) member(s) with the correct answer was not the member with the highest confidence. Respectively, for the confidence in reason: four out of 11 and six out of 11.

Performance on the transfer task was coded as correct only if a participant provided the correct answer to both problems. Thirty-four participants (39%) provided the correct answer, a significant improvement over initial performance (22%), McNemar's chi-squared: $\chi^2(1, N = 87) = 5.6$, p = .018.

Discussion. Experiment 3 replicates previous demonstrations that the "truth wins" scheme describes the outcome of the group discussion of intellective problems. Although confidence correlated with correctness of initial answer in this experiment, correctness of initial answer was still a better predictor of which position would be adopted by the group than initial confidence. In other words, in some groups a member who had found the correct answer individually was able to convince her peers who had the wrong answer despite the fact that she was not the most confident group member, or even that she was less confident than the average of the others. This means that when the most confident member did not have the right answer and that another member did, the former never managed to sway the group. A3 is better supported than C3. Also supporting the Argument Explanation (A2 in particular) of group performance, most of the participants who

[3] One group failed to find the correct answer despite having a member who had found it individually. However, this group provided a normatively defensible justification for their answer, arguing that although the logic of the task suggests a "Yes" answer, the categories "Married" and "Not married" do not exhaust the set of possible marital statuses—widowers or people with civil unions, for instance, might not clearly fit in either category. Given that these participants understood the logic of the task but failed to provide the answer usually seen as normatively correct, they were difficult to categorize and were excluded from all analyses.

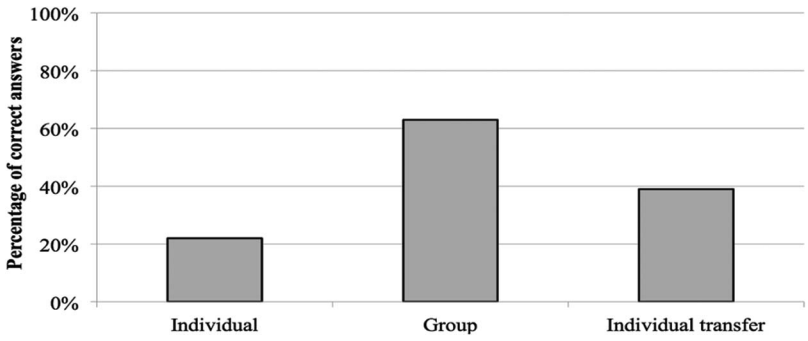

Figure 1. Percentage of correct answers in the three conditions of Experiment 3.

had been convinced to adopt the correct answer during the group discussion were able to transfer their understanding to new problems (31 out of 55).

Experiment 3 shows that even when C1 holds—confidence and correctness correlate—and that confidence can freely express itself in a face-to-face discussion, the outcome of group discussion is better explained as resulting from an exchange of arguments than from an evaluation of the member's relative confidence.

Experiment 4: Effects of Confidence and Arguments in Group Discussion Among Children

The most convincing demonstration that confidence trumps argumentation in group discussion might be the study of Levin and Druyan (1993) showing that children (sixth to 10th graders) tended to be convinced by the most confident group member, whether she was right or not, in one Piagetian task and a physics problem. The results of Experiments 1, 2, and 3, along with prior results, suggest that this tendency is typically not true of adults: Arguments, not confidence, account for the outcome of group discussion of intellective tasks. However, the Confidence Explanation might hold among children, as they might find it more difficult to justify their point of view or be less receptive to others' arguments.

Experiment 4 was similar to Experiment 3, but it was conducted among fifth graders. Instead of the disjunctive reasoning task, three intellective problems adapted to the participants were used.

Method

Participants and design. The participants were 151 French fifth graders from four different schools (six classes, 76 girls; M_{Age} = 10.8 years, SD = 0.58). The experiment was conducted during class. Participants solved three problems individually, then they solved two problems in group, and one problem was explained by the experimenter. Finally, they individually solved a transfer version of each problem.

Materials and procedure. The task consisted of three arithmetic problems with an intuitive wrong answer. Problem A was a simplified version of the "widget problem" from the Cognitive Reflection Test; Problem B was taken from a French mathematics textbook; Problem C was inspired by the "horse trader" problem from Maier and Solem (1952; see Appendix C). First, participants were asked to solve the three problems and to write their answer, providing after each a measure of confidence (six choices, ranging from "I answered completely randomly" to "Absolutely sure of my answer"), a reason for their answer, and an evaluation of this reason (six choices, ranging from "I have no justification at all for my answer" to "I think these are perfect justifications for my answer"). Then 36 groups of four to five participants were formed and asked to solve two of the three problems again after they had reached a consensus. The third problem was solved and explained by the experimenter collectively. Finally, participants had to solve a transfer task individually: three new versions of the initial problems with superficial changes. Both problem versions and order were pseudorandomized. Each problem was discussed by the groups in four classes and explained by the experimenter in two classes.

Results and Discussion

Results. As in Experiments 2 and 3 an answer was coded as correct only when an appropriate justification was provided. We observed 10 occurrences of participants giving the right answer without an appropriate justification during the pretest. As the status of these participants is unclear, especially during the group discussion, we have excluded them from the analyses along with the participants who happened to be in the same group discussion for the relevant problem (seven group occurrences, making for 33 individual occurrences).

The success rates of individuals and groups were, for Problems A, B, and C, respectively: 10 (7%), 19 (13%), and 34 (25%) individual correct answers and 30 (31% or seven groups out of 23), 59 (61% or 14 groups out of 23), and 45 (56% or 11 out of 19) correct answers after group discussion (see Figure 2).

The improvement was significant for each problem when each participant was treated as a data point for the postdiscussion answer (Fisher exact tests indicate that for each problem $p < .001$). When each group was treated as a data point for the postdiscussion answer the improvement was significant only for Problems A and B (respectively, $p = .013$, and $p = .001$) but not for Problem C ($p = .26$), suggesting that group discussion is not as efficient in the case of Problem C.

Of the group discussions in which at least one member had initially answered correctly (six groups for Problem A, nine for Problem B, and 13 for Problem C) the "truth wins" scheme applied in, respectively, six (100%), eight (89%), and eight (62%) groups, including five, seven, and four cases for which only one member had the correct answer. Respectively for each problem, of the 17,

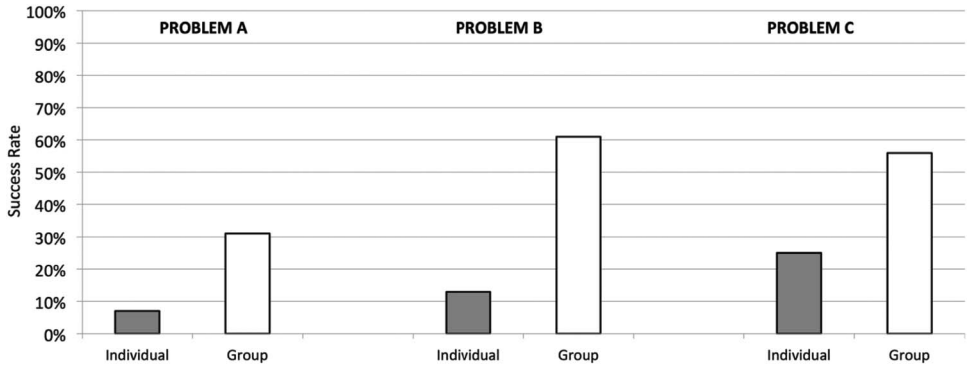

Figure 2. Success rates in individual settings and after group discussion for each problem.

14, and six groups in which no member had initially found the correct answer, one (6%), six (42%), and two (33%) groups provided the correct answer.

On the basis of these results, it appears that Problem C behaves in a different way than Problems A and B. This is coherent with previous results obtained with adults (Johnson & Torcivia, 1967), as discussed in the conclusion. Further analyses combine Problems A and B and treat Problem C separately. To analyze confidence, we separated the answers in three categories: correct answer, intuitive wrong answer, and other wrong answers.

For Problems A and B (Figure 3, left side), average confidence in answer was 80% for the right answers (*Mdn* = 80.0, *SD* = 20.7), 73% for the intuitive wrong answers (*Mdn* = 80.0, *SD* = 22.4), and 65% for the other wrong answers (*Mdn* = 60.0, *SD* = 23.3). Average confidence in reason was 78% for the right answers (*Mdn* = 80.0, *SD* = 23.5), 68% for the intuitive wrong answers (*Mdn* = 60.0, *SD* = 24.8), and 57% for the other wrong answers (*Mdn* = 60.0, *SD* = 32.1). For both measures participants who had given the intuitive wrong answer were more confident than those who had given another wrong answer (confidence in answers: $W = 5360$, $p = .010$; confidence in reason: $W = 5208$, $p = .021$). Participants with the right answer were more confident than those who had given the intuitive wrong answer, a difference nonsignificant for the confidence in answers ($W = 2072$, $p = .142$) but significant for the confidence in reason ($W = 1883$, $p = .039$).

For Problem C (Figure 3, right side, average confidence in answers for the right answer, the intuitive wrong answer and for the other wrong answers were, respectively, 84% (*Mdn* = 80.0, *SD* = 19.2), 82% (*Mdn* = 80.0, *SD* = 21.0), and 66% (*Mdn* = 60.0, *SD* = 26.7). The average confidences in reason for those three categories of answers were 75% (*Mdn* = 80.0, *SD* = 16.8), 64% (*Mdn* = 80.0, *SD* = 35.7), and 58% (*Mdn* = 60.0, *SD* = 28.3). As for Problems A and B, participants who had given the intuitive wrong answer were more confident than those with an-

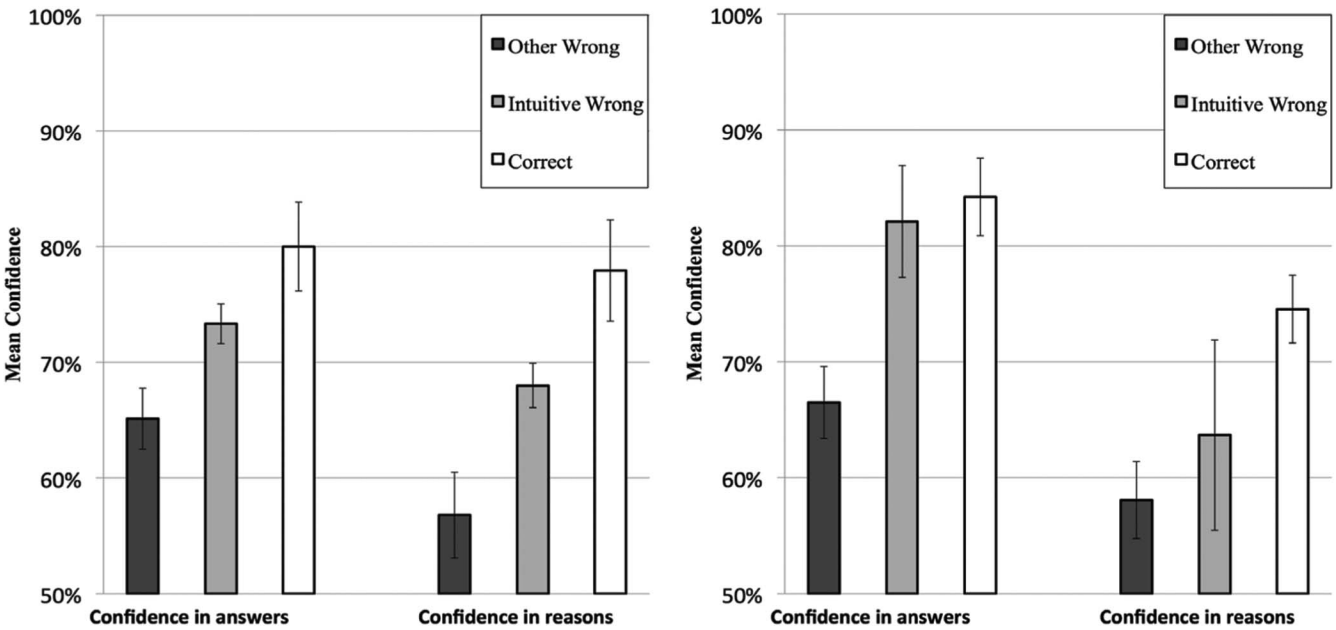

Figure 3. Mean confidence in answers and mean confidence in reasons with standard error bars for the three categories of answer in Experiment 4. On the left side for Problems A and B, on the right side for Problem C.

other wrong answer, significantly for the confidence in answer ($W = 465$, $p = .020$) but not for the confidence in reason ($p = .16$). Contrary to Problems A and B, participants who had given the right answer were not significantly more confident than those who had given the intuitive wrong answer on either measures of confidence (in answers: $W = 299$; $p = .77$; in reason: $W = 299$; $p = .77$).

The group members who had the correct answer before group discussion tended to be more confident than their peers, making it difficult to disentangle the influence of confidence and arguments on the outcome of group discussion. Still, in some groups there was a dissociation between what the group answer would be if participants followed the more confident group member (Hypothesis C3) than if they followed the group member with the correct answer (Hypothesis A3). Hypothesis C3 can take two forms: Eeither the participants adopt the answer of the single most confident group member or they adopt the answer defended by the group members who give the same answer and are on average more confident.

For Problems A and B, regarding confidence in answer, there were 12 cases in which the member(s) with the correct answer was (were) not the single most confident member(s). In 11 of these cases the group provided the right answer. There were 10 cases in which the correct answer was not given by the members with the highest average confidence. In nine of these cases the group provided the right answer. Regarding confidence in reason, there were 12 cases in which the member(s) with the correct answer was not the single most confident member. In 11 of these cases the group provided the right answer. There were nine cases in which the correct answer was not given by the members with the highest average confidence. In eight of these cases the group provided the right answer.

McNemar's chi squared tests indicate that over the 15 groups in which at least one member had the correct answer, the group outcomes were better predicted by who had the correct answer than by what answer was defended by the most confident members on average—for the confidence in answer, $\chi^2(1, N = 15) = 7.1$, $p = .008$; for the confidence in reason, $\chi^2(1, N = 15) = 6.1$, $p = .013$—or by the single most confident member—for the confidence in answer, $\chi^2(1, N = 15) = 9.1$, $p = .003$; for the confidence in reason, $\chi^2(1, N = 15) = 9.1$, $p = .003$.

For Problem C, regarding confidence in answer, there were seven cases in which the member(s) with the correct answer was (were) not the single most confident member(s). In four of these cases the group provided the right answer. There were four cases in which the correct answer was not given by the members with the highest average confidence. In one of these cases the group provided the right answer. Regarding confidence in reason, there were six cases in which the member(s) with the correct answer was not the single most confident member. In four of these cases the group provided the right answer. There were five cases in which the correct answer was not given by the members with the highest average confidence. In two of these cases the group provided the right answer.

In contrast to Problems A and B, McNemar's chi-squared tests indicate that for Problem C, over the 13 groups in which at least one member had the correct answer to the Problem C, the group outcomes were not better predicted by who had the correct answer than by what answer was defended by the most confident members on average—for the confidence in answer, $\chi^2(1, N = 13) = 0$, $p = 1$; for the confidence in reason, $\chi^2(1, N = 13) = 0$, $p = 1$—or by the single most confident member—for the confidence in answer, $\chi^2(1, N = 13) = 0.8$, $p = .37$; for the confidence in reason, $\chi^2(1, N = 13) = 0.166$, $p = .68$.

Regarding the transfer problems, there were, respectively, for each problem, 26 (57%), 27 (55%), 31 (57%) correct answers when the problem had been explained by the experimenter and 22 (23%), 37 (39%), and 31(39%) when the problem had been discussed in group. Fisher exact tests indicate that for each problem, the participants performed significantly better at the transfer task following an explanation than at the first individual resolution (ps $< .001$ for the three problems), and participants also performed significantly better at the transfer task following group discussion than at first individual resolution (A: $p < .001$, B: $p < .001$, and C: $p = .046$). Since all the participants who had been provided an explanation by the experimenter were presented with the right answer and its explanation, we compared their performance on the transfer task to that of participants who had been exposed to the right answer during the group discussion, whether they adopted it or not. With the conservative assumption that all the participants exposed their point of view in the group discussion, there were, respectively, for each problem 30, 63, and 63 occurrences of a participant being exposed to, or already having the correct answer during group discussions. Out of these occurrences, respectively, 19 (63%), 35 (56%), and 28 (44%) led to a success in the transfer task, nonsignificant differences with the success rates following the experimenter's explanation (Fischer exact tests, respectively, for each problem: $p = .64$, $p = 1$, and $p = .20$).

Discussion. Experiment 4 extends the results of Experiment 3 to younger participants: fifth graders. Even when they are in a minority, participants who have the correct answer are more likely to convince their peers than to be convinced by them. Confidence was relatively well calibrated, with participants providing the right answers being more confident (both in their answers and in their reason) than those with the wrong answer (either the intuitive wrong answer or other wrong answers). In spite of this correlation between correctness and confidence, correctness was a better predictor of group performance than confidence. Again, A3 is better supported than C3. Participants who had been exposed to the right answer during the group discussion performed well on a transfer problem. Remarkably, their performance was similar to that of participants who had been provided with the right answer and an explanation by the experimenter, demonstrating the power of peer arguments—even when compared with a respected epistemic authority—to yield genuine understanding of the tasks, in support of A2. Experiment 4 demonstrates that even in 10-year-olds, presumably less skilled at argumentation than college students, argument quality and not confidence is the best explanation of group performance.

An important qualifier to this conclusion is that it applied chiefly to two of the three problems used. In the last problem (C, the horse trader), confidence played a much more important role. We offer an explanation for this finding in the conclusion, as well as an interpretation of the difference between the present results and those obtained by Levin and Druyan (1993).

Conclusion

To know how much the exchange of arguments or the recognition of confidence drives the performance of groups discussing intellective tasks, we tested predictions made by an Argument Explanation (A1, A2, and A3) and by a Confidence Explanation (C1, C2, and C3). Supporting A1 (*good arguments change people's mind*), Experiments 1 and 2, in which participants were exposed to arguments against their initial answer, showed that a single good argument, with no confidence marker, can change people's mind on an intellective task, whereas an argument for the wrong answer did not make any participant who had provided the right answer change their minds.

Experiments 3 (with adults) and 4 (with fifth graders) asked participants to solve intellective tasks individually, in groups, and then to solve a transfer problem. Supporting A2 (*people are sometimes able to transfer the understanding they have gained during group discussion to other tasks*), these experiments showed that participants who have been exposed to the correct answer during group discussion were more likely to solve the transfer problem than they had been to solve the initial problem on their own. Supporting A3 (*being right is a better predictor of one's ability to convince one's peers than being confident*) over C3 (*being confident is a better predictor of one's ability to convince one's peers than being right*), in both experiments some groups contained a member who had initially found the correct answer but who wasn't the most confident, or group members who had found the correct answer but who weren't, on average, the most confident. Across Experiments 3 and 4, using both means to make predictions based on confidence, and using both confidence in answer and confidence in reason, in at least 80% of the cases the group adopted the answer of the correct member(s) even when they were not the most confident (except for Problem C of Experiment 4, discussed below).

Previous measures of confidence in reasoning tasks have yielded contrasted results, some reporting a correlation between confidence and correctness (De Neys, Cromheeke, & Osman, 2011) others not (Shynkaruk & Thompson, 2006). On the whole, the present results support C1 (*participants with the good answer are more confident*) for the intellective problems studied, whether we examine confidence in the correctness of the answers or confidence in the reason for the answer. Still, with regards to the Confidence Explanation it should be stressed that the levels of confidence in the wrong answers are remarkably high, making it less likely that the participants with the correct answer might sway others thanks to a large confidence gap. For instance, in Experiment 2a, for the participants with the intuitive wrong answer, the modal confidence estimate (35% of the participants) was the highest point of the confidence scale: "As confident as in the things I'm most confident about."

What might be the most surprising finding emerging from the present experiments is that in Experiment 2, C2 (*the most confident participants are the least likely to change their minds*) did not hold. Participants' initial confidence in their answers did not predict their susceptibility to accept a contrary argument and change their minds. Even more surprisingly, among the participants who had given the intuitive but wrong answer, those who were the most confident in their reason were also the most likely to change their minds. The lack of a positive influence of confidence on susceptibility to accept a contrary argument suggests that the reasoning mechanisms evaluating the argument can act independently of the strength of the belief in the position attacked by the argument (as predicted by the argumentative theory of reasoning Mercier & Sperber, 2011). This result seemingly runs against many previous observations of the influence of prior beliefs on argument evaluation (e.g., Lord, Ross, & Lepper, 1979). The argument for the good answer used in Experiment 2 might explain this discrepancy. This argument was simple enough that it could have provoked, among some participants, an "ahah!" experience such that they were immediately persuaded by the argument and did not start producing counterarguments influenced by their previous beliefs. This experience would be similar to other types of Eureka experiences, except that it bears on an argument rather than an answer (for an example of eureka experience in group settings, see Shaw, 1932). To test this explanation, the effect will have to be replicated and extended to other arguments and other beliefs.

These results, in addition to the literature mentioned in the introduction, make it clear that arguments, rather than confidence, are the main factor explaining the performance of groups discussing intellective tasks. How are we to interpret, then, the apparent exceptions to the effectiveness of arguments in intellective tasks? Stanovich and West (1999) observed that for some problems, such as the base rate problems, arguments for the normative answer were not more convincing than arguments for the nonnormative answer. They noted, however, that the normative answer to these problems has long been and, for some problems, is still discussed in the literature. If the community of experts has had great difficulty in agreeing on the correct answer to these problems, it is hardly surprising that untutored participants also fail to find the arguments for one side more convincing than those for the other side. Similarly, the problem for which Levin and Druyan (1993) found that group discussion led to poorer performance was particularly difficult: It implicitly asked 13-year-olds to understand the difference between standard and angular speed, concepts with which they were presumably not acquainted. It can be argued that in this case, providing the wrong answer reflected a better grasp of the principles of physics involved than providing the "right" answer—throughout history, wrong physical theories have been defended by very good reasons (see, e.g., Bozzi, 1958). Accordingly, although the groups ended up with more wrong answers, it might still be the arguments, rather than the confidence, of the wrong group members that convinced their peers—and these arguments might have been quite sensible. To sum up, if no one in the population is able to reach a genuine understanding of the problem because they are lacking the relevant concepts, groups might not perform better than individuals. Participants cannot be expected to discover on their own Bayes' theorem or the distinction between standard and angular speed. While it is interesting to probe what happens during group discussion when genuine understanding is not accessible, the outcome can hardly be an indictment of our abilities to recognize good arguments and be swayed by them. This raises the issue of distinguishing arguments that promote understanding when the concepts involved are all sufficiently mastered by the individuals, from arguments that rely on concepts that are yet out of reach of the participants and open the path to sensible arguments defending the wrong answer.

The problem of the stamp collector (adapted from the "horse trader" as Problem C of Experiment 4) offers a seemingly different case, as it is clear from the justifications that the participants who provided the right answer did so for the right reasons, such as simply adding and subtracting the transaction amounts in turn (see Appendix C for the problem). Yet they were not able to convince their peers with the same regularity as for the other problems. Indeed, the most confident members often convinced the group, even if they had the

wrong answer. This result—with fifth graders—replicates the findings Johnson and Torcivia (1967) obtained with adult dyads. A possible explanation is that although some participants can understand the task, they might find it difficult to contradict a confident member advocating the wrong intuitive answer because they might find it challenging to explain why the wrong answer is wrong. In fact, a possibility that needs further testing is that the superficial features of the stamp collector problem induce a misleading semantic structure (Bassok, Wu, & Olseth, 1995), a "win and lose" structure that is evoked because this is the same object that is bought and sold back and forth; each transaction seems symmetrical and to generate a gain if it has been bought cheaper than the price it was sold and a lost in the reverse case. According to this "win and lose" structure, the stamp collector (in this version) loses 10€ when he buys back 80€ a stamp that he previously sold 70€. It seems to cancel his first 10€ gain and then the second 10€ gain seems to constitute the whole profit. This interpretation is very appealing and hard to inhibit because there are many seemingly analogous cases in which it is relevant such as when someone wins and loses money: If I lose 60€ (seemingly analogous to buying the stamp) and then I win 70€ (seemingly analogous to selling the stamp), and then I lose 80€ and then win 90€, it is true that there will remain a positive balance of 10€. The specificity of the stamp collector problem is that buying is not losing; it works as two independent transactions, each one generating its own profit. As a matter of fact, the wrong intuitive solution decreases radically when it is not the same object, but two distinct objects that are considered (a decrease from 46% to 12% of the wrong intuitive answer among undergraduate students) because the " win and lose" structure is no more evoked when two distinct objects are involved (Sander & Mathieu, 2005); in the latter case the misleading structure does not compete anymore with the relevant interpretation, which corresponds to usual "buy and sales" scenario in which a person buys and sells several objects and in which the total gain is the sum of each individual gain. This way of thinking is masked by the "win and lose" structure that constitutes the interpretation framework of the situation when the object is the same, and even participants who use the relevant algorithm of adding and subtracting gains and losses might not grasp the reason why "win and lose" is fallacious in this context.

Our results demonstrate that under some conditions at least, participants are able to produce and evaluate arguments in such a way that those who have the correct answer to an intellective problem convince their peers with near perfect accuracy, even when they are not the most confident (across Experiment 3, and Problems A and B of Experiment 4, only one group did not follow the "truth wins" scheme). This is an important demonstration of sound argumentative competence, in line with the predictions of, for instance, the argumentative theory of reasoning (Mercier & Sperber, 2011). However, it is clear as well that the conditions of demonstrability have to be well respected for this to be the case. As noted in the original definition, "the group members who are not themselves able to solve the problem must have sufficient knowledge of the system to recognize and accept a correct solution if it is proposed by another group member, [and] the correct member must have sufficient ability, motivation, and time to demonstrate the correct solution to the incorrect members" (Laughlin & Ellis, 1986, p. 180). This excludes several of the problems for which arguments for the correct answer had failed to convince participants (e.g., Levin & Druyan, 1993; Stanovich & West, 1999). In order to evaluate the role played by different psychological mechanisms in group reasoning, it is critical to rely on problems with well-specified properties.

References

Anderson, C., & Kilduff, G. J. (2009). Why do dominant personalities attain influence in face-to-face groups? The competence-signaling effects of trait dominance. *Journal of Personality and Social Psychology, 96,* 491–503. doi:10.1037/a0014201

Aramovich, N. P., & Larson, J. R. (2013). Strategic demonstration of problem solutions by groups: The effects of member preferences, confidence, and learning goals. *Organizational Behavior and Human Decision Processes, 122,* 36–52. doi:10.1016/j.obhdp.2013.04.001

Bassok, M., Wu, L., & Olseth, K. L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory & Cognition, 23,* 354–367. doi:10.3758/BF03197236

Bozzi, P. (1958). Analisi fenomenologica del moto pendolare armonico [Phenomenological analysis of pendular harmonic motion]. *Rivista Di Psicologia, 52,* 281–302.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, England: MIT Press.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS One, 6,* e15954. doi:10.1371/journal.pone.0015954

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*. Advance online publication. doi:10.3758/s13423-013-0384-5

Doise, W., & Mugny, G. (1984). *The social development of the intellect*. Oxford, England: Pergamon Press.

Evans, J. S. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin, 128,* 978–996. doi:10.1037/0033-2909.128.6.978

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19,* 25–42. doi:10.1257/089533005775196732

Gamo, S., Sander, E., & Richard, J.-F. (2010). Transfer of strategy use by semantic recoding in arithmetic problem solving. *Learning and Instruction, 20,* 400–410. doi:10.1016/j.learninstruc.2009.04.001

Johnson, H. H., & Torcivia, J. M. (1967). Group and individual performance on a single-stage task as a function of distribution of individual performance. *Journal of Experimental Social Psychology, 3,* 266–273. doi:10.1016/0022-1031(67)90028-5

Koriat, A. (2012). When are two heads better than one and why? *Science, 336,* 360–362. doi:10.1126/science.1216549

Kuhn, D., & Lao, J. (1996). Effects of evidence on attitudes: Is polarization the norm? *Psychological Science, 7,* 115–120. doi:10.1111/j.1467-9280.1996.tb00340.x

Laughlin, P. R. (2011). *Group problem solving*. Princeton, NJ: 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. doi:10.1016/0022-1031(86)90022-3

Levesque, H. J. (1986). Making believers out of computers. *Artificial Intelligence, 30,* 81–108. doi:10.1016/0004-3702(86)90068-8

Levin, I., & Druyan, S. (1993). When sociocognitive transaction among peers fails: The case of misconceptions in science. *Child Development, 64,* 1571–1591. doi:10.2307/1131553

Littlepage, G. E., Schmidt, G. W., Whisler, E. W., & Frost, A. G. (1995). An input-process-output analysis of influence and performance in problem-solving groups. *Journal of Personality and Social Psychology, 69,* 877–889. doi:10.1037/0022-3514.69.5.877

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37,* 2098–2109. doi:10.1037/0022-3514.37.11.2098

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,* 854–870. doi:10.1037/0022-3514.92.5.854

Maier, N. R., & Solem, A. R. (1952). The contribution of a discussion leader to the quality of group thinking: The effective use of minority opinions. *Human Relations, 5,* 277–288. doi:10.1177/001872675200500303

Mercier, H., Deguchi, M., Van der Henst, J.-B., & Yama, H. (2014). *The benefits of argumentation are cross-culturally robust: The case of Japan*. Manuscript submitted for publication.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34,* 57–74. doi:10.1017/S0140525X10000968

Miller, S. A., & Brownell, C. A. (1975). Peers, persuasion, and Piaget: Dyadic interaction between conservers and nonconservers. *Child Development, 46,* 992–997.

Moshman, D., & Geil, M. (1998). Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning, 4,* 231–248. doi:10.1080/135467898394148

Nussbaum, E. M. (2008). Collaborative discourse, argumentation, and learning: Preface and literature review. *Contemporary Educational Psychology, 33,* 345–359.

Perret-Clermont, A.-N. (1980). *Social interaction and cognitive development in children*. London, England: Academic Press.

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making, 17,* 39–57. doi:10.1002/bdm.460

Sander, E., & Mathieu, N. (2005). *Individualization as influencing semantic alignment in mathematical word problem solving*. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the Cognitive Science Society* (p. 1628). Mahwah, NJ: Erlbaum.

Sander, E., & Richard, J.-F. (1997). Analogical transfer as guided by an abstraction process: The case of learning by doing in text editing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1459–1483. doi:10.1037/0278-7393.23.6.1459

Shaw, M. E. (1932). A comparison of individuals and small groups in the rational solution of complex problems. *The American Journal of Psychology, 44,* 491–504. doi:10.2307/1415351

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition, 34,* 619–632. doi:10.3758/BF03193584

Slovic, P., & Tversky, A. (1974). Who accepts Savage's axiom? *Behavioral Science, 19,* 368–373. doi:10.1002/bs.3830190603

Stanovich, K. E., & West, R. F. (1999). Discrepancies between normative and descriptive models of decision making and the understanding/acceptance principle. *Cognitive Psychology, 38,* 349–385. doi:10.1006/cogp.1998.0700

Toplak, M. E., & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 94,* 197–209. doi:10.1037/0022-0663.94.1.197

Trognon, A. (1993). How does the process of interaction work when two interlocutors try to resolve a logical problem? *Cognition and Instruction, 11,* 325–345. doi:10.1080/07370008.1993.9649028

Tudge, J. (1989). When collaboration leads to regression: Some negative consequences of socio-cognitive conflict. *European Journal of Social Psychology, 19,* 123–138. doi:10.1002/ejsp.2420190204

Van Swol, L. M., & Sniezek, J. A. (2005). Factors affecting the acceptance of expert advice. *British Journal of Social Psychology, 44,* 443–461. doi:10.1348/014466604X17092

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes, 69,* 237–249. doi:10.1006/obhd.1997.2685

Zarnoth, P., & Sniezek, J. A. (1997). The social influence of confidence in group decision making. *Journal of Experimental Social Psychology, 33,* 345–366. doi:10.1006/jesp.1997.1326

Appendix A

Confidence Scales in Experiment 2

Confidence in Answer Scale

- Not confident at all
- A little confident
- Somewhat confident
- Quite confident
- Very confident
- Extremely confident
- Perfectly confident
- So confident I can't imagine ever changing my mind
- As confident as the things I'm most confident about

Confidence in Reason Scale

- Very poor reason
- Poor reason
- Average reason
- Good reason
- Very good reason
- Extremely good reason
- A demonstrative reason that perfectly supports its conclusion

(*Appendices continue*)

Appendix B

Transfer Problems in Experiment 3

Transfer Problem 1

In a building, the person living on the 2nd floor does not have a dog and the one living on the 4th floor does have a dog. In this building, is someone who does not have a dog living just below someone who does have a dog?

Yes, No, Cannot be determined

Transfer Problem 2

On the marketplace, there are clients and sellers. Among them are Michel, Marie, and Philippe. Michel is looking at Marie, and Marie is looking at Philippe. Michel sells fish and Philippe sells vegetables. Is there, among those three people, a seller who is looking at a client?

Yes, No, Cannot be determined

Appendix C

Problems Used in Experiment 4

Problem A—First Version

Some birds are flying around a mountain. It takes 30 minutes for a group of 3 birds to circle the mountain. How long will a group of 4 birds take to circle the mountain?

Correct answer: 30

Wrong intuitive answer: 40

Problem A—Second Version

In a tall building, the lift is broken and people have to use the stairs. In 1 hour, a group of 4 people climbs 40 stairs. In 1 hour, how many stairs will a group of 8 people climb?

Correct answer: 40

Wrong intuitive answer: 20 or 80

Problem B—First Version

Julie puts her books on a shelf in her bedroom. She realizes that her favorite book is the 33rd from the left and the 44th from the right. How many books does Julie have on her shelf?

Correct answer: 76

Wrong intuitive answer: 77

Problem B—Second Version

Marc is counting how many houses there are in his street. His house is the 62nd from the bottom of the street and the 34th from the top of the street. How many houses are there in Marc's street?

Correct answer: 95

Wrong intuitive answer: 96

Problem C—First Version

A stamp collector bought a stamp 60€ and sold it 70€. Then he bought back the stamp for 80€ and sold it for 90€. After all this, how much money did he win?

Correct answer: 20

Wrong intuitive answer: 10

Problem C—Second Version

A plane is flying at an unknown altitude. The plane goes down 400 meters and then up 500 meters. Then it goes down again 600 meters and back 700 meters. In the end, how much higher is the plane's altitude compared to his starting altitude?

Correct answer: 200

Wrong intuitive answer: 100

In the group discussion condition, whether the first or the second version was used as initial problem and as transfer was counterbalanced.

Received November 26, 2013
Revision received April 8, 2014
Accepted April 17, 2014 ■