Journal of Experimental Psychology: General

Argumentation and the Diffusion of Counter-Intuitive Beliefs

Nicolas Claidière, Emmanuel Trouche, and Hugo Mercier

Online First Publication, May 11, 2017. http://dx.doi.org/10.1037/xge0000323

CITATION

Claidière, N., Trouche, E., & Mercier, H. (2017, May 11). Argumentation and the Diffusion of Counter-Intuitive Beliefs. *Journal of Experimental Psychology: General*. Advance online publication. http://dx.doi.org/10.1037/xge0000323

Journal of Experimental Psychology: General

© 2017 American Psychological Association
0096-3445/17/$12.00 http://dx.doi.org/10.1037/xge0000323

Argumentation and the Diffusion of Counter-Intuitive Beliefs

Nicolas Claidière
Centre National de la Recherche Scientifique, France, and
University Aix-Marseille

Emmanuel Trouche and Hugo Mercier
Centre National de la Recherche Scientifique, France, and
University Lyon 2

Research in cultural evolution has focused on the spread of intuitive or minimally counterintuitive beliefs. However, some very counterintuitive beliefs can also spread successfully, at least in some communities—scientific theories being the most prominent example. We suggest that argumentation could be an important factor in the spread of some very counterintuitive beliefs. A first experiment demonstrates that argumentation enables the spread of the counterintuitive answer to a reasoning problem in large discussion groups, whereas this spread is limited or absent when participants can show their answers to each other but cannot discuss. A series of experiments using the technique of repeated transmission show that, in the case of the counterintuitive belief studied: (a) arguments can help spread this belief without loss; (b) conformist bias does not help spread this belief; and (c) authority or prestige bias play a minimal role in helping spread this belief. Thus, argumentation seems to be necessary and sufficient for the spread of some counterintuitive beliefs.

Keywords: counterintuitive beliefs, argumentation, cultural evolution, conformist bias, prestige bias

Supplemental materials: http://dx.doi.org/10.1037/xge0000323.supp

Some ideas have managed to spread in human societies despite being highly counterintuitive, such as heliocentrism. Remarkably, these ideas have spread in the face of beliefs that were not only more intuitive, but also more widespread and held by the most prestigious members of the relevant group, raising an interesting challenge for the study of cultural evolution.

Studies have shown that cultural evolution often converges on the most intuitive version of a given cultural product. Intuitiveness has been operationalized in many different ways, but it can be broadly defined as a property of cultural products that makes them easy to process. For instance, Miton, Claidière, and Mercier (2015) conducted a series of experiments with stories involving bloodletting—the practice of drawing blood to cure an ailment. The stories in which bloodletting was described in a way predicted to be more intuitive—for instance, when the cut was performed on the ailing body part—were better remembered, and less intuitive stories tended to converge toward more intuitive versions. Convergence toward intuitive versions of cultural products has been demonstrated in various areas of human culture (language: Griffiths, Kalish, & Lewandowsky, 2008; Kirby, Cornish, & Smith, 2008; Reali & Griffiths, 2009; medicine: Miton, Claidière, & Mercier, 2015; art: Morin, 2013), and in other animals (visual stimuli: Claidière, Smith, Kirby, & Fagot, 2014; bird song: Feher, Wang, Saar, Mitra, & Tchernichovski, 2009; foraging strategy: Laland & Williams, 1997).

Theoretical analyses have also revealed the origin and strength of this result: when cultural products (artifacts, behaviors, or representations) spread in a population, transformations of these products that occur during transmission progressively accumulate and tend to be directed by pre-existing biases, toward easier to process versions of these products (Claidière & Sperber, 2010; Griffiths et al., 2008; Kalish, Griffiths, & Lewandowsky, 2007; Kirby, Dowman, & Griffiths, 2007). Given the strength and generality of these results it is intriguing that counterintuitive beliefs can sometimes dislodge and replace more intuitive alternatives.

One possible explanation relies on the properties of counterintuitive beliefs: counterintuitive ideas can become attractive in virtue of their counterintuitiveness. For instance, many beliefs in religious and other supernatural entities are counterintuitive: a ghost has the counterintuitive property of being invisible, say. Boyer (2001) has suggested that in fact such beliefs are ideal for cultural transmission because they are *minimally* counterintuitive: a ghost is invisible but has the mind of a human being and is, therefore, both easy to understand (since it is mostly intuitive), and memorable (because of the counterintuitive property). Many experiments have shown that beliefs such as ghosts are better re-

Nicolas Claidière, Nicolas Claidière, Laboratoire de Psychologie Cognitive, Centre National de la Recherche Scientifique (UMR7290), France, and University Aix-Marseille; Emmanuel Trouche and Hugo Mercier, Institut des Sciences Cognitives Marc Jeannerod, Centre National de la Recherche Scientifique (UMR5304), France, and University Lyon 2.

NC gratefully acknowledges financial support from the ASCE program (ANR-13-PDOC-0004) and the LICORNES program (ANR-12-CULT-0002) of the Agence Nationale de la Recherche, Emmanuel Trouche from the Direction Générale de l'Armement, and HM from the Swiss National Science Foundation (Ambizione Grant PZ00P1_142388/1) and the Agence Nationale de la Recherche (Tremplin-ERC Grant ANR-16-TERC-0001–01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. None of the present results have been published elsewhere.

Correspondence concerning this article should be addressed to Hugo Mercier, Institut des Sciences Cognitives Marc Jeannerod, 67 boulevard Pinel, 69675 Bron, France. E-mail: hugo.mercier@gmail.com

membered than either purely intuitive, or much less intuitive alternatives (for reviews, see Barrett, 2000; Boyer, 2001).

According to this explanation, the cultural evolution of minimally counterintuitive beliefs follows from the general principles highlighted previously: the more appealing beliefs—overall—tend to spread. However, many beliefs that are much more than minimally counterintuitive have spread and remained stable—for instance, that the earth revolves around the sun. To highlight the contrast between these beliefs and minimally counterintuitive beliefs, we will call them *very counterintuitive beliefs*. For instance, the theologically correct Christian belief in an omnipotent and omnipresent God violates many intuitions. An interesting find was that even if this belief is apparently widespread, those who believe in it tend to generate inferences in line with an only minimally counterintuitive version of this God—for instance, one that can only attend to one prayer at a time (Barrett & Keil, 1996). We can surmise that heliocentrism is another such very counterintuitive belief, because it contradicts strong intuitions—mainly, that we see the sun move and that we do not feel the earth moving—without being immediately consistent with any intuition.

A possible explanation of the spread of very counterintuitive beliefs relies not on the intrinsic properties of the beliefs themselves, but on their source. For instance, if a prestigious individual adopts a very counterintuitive belief, this belief could then be adopted by other members of the population who imitate prestigious individuals (Boyd & Richerson, 1985; Henrich & Gil-White, 2001; Richerson & Boyd, 2005). The spread of this belief could then be reinforced and stabilized by a conformist tendency (i.e., the adoption of beliefs held by a majority of individuals; Boyd & Richerson, 1985; Henrich & Boyd, 1998; Richerson & Boyd, 2005). Boyd, Richerson, and Henrich in particular have argued that such processes can lead in some cases to the spread and stabilization of any cultural product, including maladaptive or counterintuitive ones (see Henrich, 2015; Mercier, in press; Richerson & Boyd, 2005). Thus, one could imagine that a combination of prestige bias and conformist bias could account for the spread and stability of very counterintuitive beliefs. It is indeed plausible that these factors might play an important role in the adoption of very counterintuitive beliefs. For instance, most people nowadays believe in scientific theories through trust in teachers and scientists (although this is likely not before prestige per se, but to sensible deference to epistemic authority), and possibly conformity.

A potential difficulty with an explanation in terms of prestige and conformity is that some very counterintuitive started spreading despite being defended by individuals who were in a minority and were not particularly prestigious. On the contrary, it is only after the counterintuitive beliefs had been accepted by the community and their value recognized that their creators were endowed with prestige. Einstein, Galileo, or Newton are good examples, but that is true of just about any influential scientist, and also, to some extent, of theologians and philosophers, who also spread very counterintuitive beliefs.

An interesting property of a significant class of very counterintuitive beliefs is that they follow from deductive reasoning. Deductive demonstrations have played a significant role in human culture at least since ancient Greek philosophers developed formal models of proofs in logic and mathematics. These methods allowed later scholars to develop very counterintuitive beliefs—for instance non-Euclidian geometries. In turn these very counterintuitive mathematical results helped give rise to even more momentous very counterintuitive scientific theories—for instance relativity theory.

We propose that argumentation plays a crucial role in the propagation of—at least—such very counterintuitive beliefs. To study in the laboratory the effects of argumentation on the spread of very counterintuitive beliefs, we rely on problems that have a deductively valid answer that is accessible by all participants and yet found spontaneously by only a few. More specifically, most participants provide instead an intuitive but incorrect answer. For instance, consider the Bat and Ball problem (Frederick, 2005):

A bat and a ball cost $1.10 together. The bat costs $1 more than the ball. How much does the ball cost?

Studies have shown that a majority of participants provide the intuitive but incorrect answer of 10c, when the correct but counterintuitive answer is 5c (5c for the ball plus $1.05 for the bat makes $1.10 in total). Typically, participants who give the intuitive but wrong answer to such problems are very confident in their answer to begin with (De Neys, Rossi, & Houdé, 2013; Mata, Ferreira, & Sherman, 2013; Trouche, Sander, & Mercier, 2014). Therefore, these problems are ideal to study the spread of counterintuitive beliefs in the laboratory because they capture the essence of a counterintuitive problem, yet are accessible to typical participants (by contrast with most counterintuitive scientific theories).

For the problems we will use here, in which the correct answer logically follows from knowledge the participants agree on, a participant who defends the correct answer typically convinces the members of a small group (e.g., $N = 4$) to accept it (Laughlin, 2011; Trouche et al., 2014). At least two elements suggest that argumentation plays a crucial role in the transmission of the correct but counterintuitive answer in small groups. First, the transcripts show participants exchanging arguments and being convinced only when good arguments are offered (Moshman & Geil, 1998; Trognon, 1993). Second, other factors such as support from other individuals or confidence seem to play a minimal role since the correct response spreads even when it is defended by a single individual facing a unanimous group, and even if this individual is less confident than the other members (Trouche et al., 2014).

If these findings suggest that argumentation can spread counterintuitive beliefs, several questions remain unaddressed. First, to demonstrate that argumentation can spread beliefs in a large population it is necessary to show that participants who have been convinced to adopt the correct answer can themselves convince others, who can convince others, and so on. By contrast, work on cultural transmission typically shows the progressive erosion of information as it goes through repeated transmissions (e.g., Bartlett, 1932; Maxwell, 1936; Mesoudi & Whiten, 2004; Northway, 1936; Scott-Phillips, in press). This erosion may act against the spread of arguments and one might predict that across several generation of transmission arguments become less and less elaborate or precise and, therefore, less and less convincing. For instance, the arguments used by Bartlett lost nearly all of their content after a few transmission episodes (Bartlett, 1932).

Second, the ease with which argumentation can help very counterintuitive beliefs spread also depends on how critical discussion is, versus being confronted with an argument (or a series of arguments) without being able to interact with their source. If discussion were critical, then most media (printed media, TV, much of the Internet, etc.) would prove unable to transmit very counterintuitive beliefs. Third, a counterintuitive belief spreads in replacement of a more common belief and, therefore, has to overcome source-based biases such as prestige and conformist biases—that is not necessarily, or as much, the case in small groups.

The following experiments seek to establish: (a) That argumentation can enable the spread of counterintuitive beliefs in large groups and across several generations without erosion (Study 1); (b) That being exposed to a single argument, instead of a full-blown argumentative discussion, can allow the spread of counterintuitive beliefs (Studies 2a and 2b); and (c) That other factors related to the source of the belief—how many people hold it, and how authoritative is the source holding it—are less efficient than argumentation in spreading counterintuitive beliefs (Studies 3a and 3b).

Materials for All Studies

In all studies we rely on two problems, the Bat and Ball described above, and the Paul and Linda problem:

> Paul is looking at Linda and Linda is looking at John.
>
> Paul is married but John is not married.
>
> Is a person who is married looking at a person who is not married?
>
> Yes, someone who is married is looking at someone who is not married
>
> No, no one who is married is looking at someone who is not married
>
> Cannot be determined whether someone who is married is looking at someone who is not married

In this problem as well, the majority of participants provide the incorrect answer “Cannot be determined,” when the correct but counterintuitive answer is “Yes, someone who is married is looking at someone who is not married” (because Linda has to be either married or not married, and that the statement is true in both cases; Toplak & Stanovich, 2002).

The Bat and Ball, and Paul and Linda are intellective problems in the sense that the participants can understand the correct answer on the basis of the information provided and their prior understanding of mathematics (Bat and Ball) and logic (Paul and Linda; see Laughlin & Ellis, 1986). We chose these problems because they are well studied reasoning problems that have been shown to have an intuitive but wrong answer given by most participants.

Study 1: Diffusion of Counterintuitive Beliefs in Large Groups Through Discussion

This experiment extends to larger groups the previous studies showing that argumentation enables the diffusion of counterintuitive beliefs in small groups. For counterintuitive beliefs to spread in large groups, participants who do not find the correct answer on their own, and who are then convinced to accept it, must be able to convince others in turn. To establish that this is the case, we tested much larger groups than the groups tested in the group decision making literature (mean group size = 18.8 participants). Moreover, we kept track of the diffusion of the answers by asking participants to provide answers at regular time intervals throughout the experiment.

To show that it is argumentation that explains the spread of counterintuitive beliefs we use, as a control condition, groups in which participants could only show their response to others but could not discuss them (see Minson, Liberman, & Ross, 2011; Rahwan, Krasnoshtan, Shariff, & Bonnefon, 2014; Rowe & Wright, 1996).

Participants

There were 226 first year students at the University of Lyon who were recruited (71 women, $M_{\text{Age}} = 19.4$, $SD = 2.1$). They were distributed based on their class assignment to 12 groups of varying sizes ($Max = 25$, $Min = 11$, $Mean = 19$). This was the first course of the year, for first year psychology students, so we can assume that most students did not know each other before the experiment.

Materials

Participants completed the Bat and Ball problem and the Paul and Linda problem in counterbalanced order. For the Bat and Ball, the answer format was open ended. For the Paul and Linda problem, the participants had to choose one of the three possible answers. In both cases participants had to indicate their confidence in their answer on a confidence scale going from 0 to 10 (the results from this question will not be presented in detail here, but the data are available in the electronic supplementary material).

Design and Procedure

Six groups took part in the Discuss condition, and six in the Silence condition. Both conditions had two phases: an Individual phase and a Social phase. The Individual phase, presently described, was identical across the two conditions.

Individual phase. After agreeing to take part in the experiment, the participants were made to sit so that the seating arrangement could best approximate a rectangle with no empty seats. Answer sheets were distributed that contained 25 identical rows, one row for each time step, with the space for an answer to the problem and a confidence scale. After a brief explanation of the first phase of the experiment, the experiment started. The problem was displayed on the screen so that all participants could start completing it at the same time. After 20 s, the participants provided their first answer and confidence rating. Four more answers and confidence ratings were gathered at succeeding 1-min intervals.

Social phase. In the Discuss condition, participants were told that they would now be able to discuss their answers with their neighbors. Neighbors were defined as the eight (maximum) students surrounding them. Participants were told

> The goal is to reach a consensus for the whole group. So after you have made sure that you agreed with some of your neighbors, you should turn to your other neighbors to make sure that they also agree on the same answer.

After they were given the signal to start discussing, the participants had to write down their answer and confidence rating every minute. After 5 min the participants were asked every other minute if at least one of them had changed their mind. The experiment was stopped when no one had changed their mind. For this reason the length of the experiment varied both within and between conditions (from 8 to 23 measures). Time was kept by the experimenter who required everyone to write down their answer every minute. The instructions were identical in the Silence condition, except that participants were instructed only to look at the answers of their neighbors, and were prohibited from talking or writing anything besides their answers.

Statistics. We analyzed the results using Generalized Linear Mixed Models (GLMM) and followed the procedure recommended by Zuur, Ieno, Walker, Saveliev, and Smith (2009). The dependent variable was the number of correct answers in relation to the total number of participants of each group at each time step (binomial variable). We included Group as a random variable with a random intercept and a random slope depending on time to account for repeated measurements. To best compare the results between the Individual and the Social phase we limited our analysis to five measures in each phase.

Based on the design of the experiments we chose to include three explanatory variables. The first variable represented the phase of the experiment, either Individual or Social. The second variable represented the experimental condition, either Discuss or Silence. Finally, we used a time variable, representing the succession of the different measurements. To facilitate interpretation of the models' coefficients we subtracted 5.5 from the time variable, so that the intercept of the model falls between the individual and the social phases. Accordingly, the intercept of the condition variable tells us whether there is a difference between conditions at the switch from the individual phase to the social phase. Further, the intercept of Phase corresponds to the change when we switch from one phase to the other. The full details of the models and the data are available in the ESM.

We used the R software and the package lme4 to build logistic regression models with a logit link function. We analyzed the results of the two problems separately and present the results of a single model that includes the predicted three-way interaction between the three explanatory variables. We conducted only planned comparisons in relation to the hypotheses formulated based on the literature and, therefore, set α at 5%, without corrections for multiple comparisons.

Results

As predicted, we found a significant three-way interaction for both problems (Bat and Ball: GLMM, $\chi^2(df = 1) = 10.6$, $p = .001$; Paul and Linda: GLMM, $\chi^2(df = 1) = 35.2$, $p < .001$; see Table 1 and Figure 1).

Table 1
GLMM Parameters for the Bat and Ball Problem

Analysis of deviance table	χ^2	*df*	*p*-value
Phase	4.30	1.00	.04
Condition	2.72	1.00	.10
Time	36.21	1.00	.00
Phase:Condition	8.00	1.00	.00
Phase:Time	.45	1.00	.50
Condition:Time	3.97	1.00	.05
Phase:Condition:Time	10.64	1.00	.00

Random effects	Variance	*SD*
Group	.97	.99
Group:Time	.02	.14

Fixed effects	Estimate	*SE*	*z*	*p*-value
Intercept	−.34	.44	−.76	.45
Phase (Social)	−.02	.25	−.07	.95
Condition (Discuss)	.41	.63	.64	.52
Time	.34	.09	3.74	.00
Phase (Social):Condition (Discuss)	.95	.44	2.17	.03
Phase (Social):Time	−.19	.09	−2.05	.04
Condition (Discuss):Time	.06	.13	.44	.66
Phase (Social):Condition (Discuss):Time	.69	.21	3.26	.00

Note. GLMM = Generalized Linear Mixed Models.

Regarding the Bat and Ball problem, during the Individual phase, we found no difference between conditions in either intercept (Wald test, β[silence]- β[discuss] = 0.41, *SE* = 0.63, *Z* = 0.64, *p* = .52) or slope (Wald test, β[silence]- β[discuss] = 0.06, *SE* = 0.13, *Z* = 0.44, *p* = .66). In both conditions the odds of success significantly increased over time (in the Silence Condition Wald test, β[time] = 0.34, *SE* = 0.09, *Z* = 3.74, $p < .001$; in the Discuss Condition Wald test, β[time] = 0.40, *SE* = 0.10, *Z* = 4.14, $p < .001$). In the Silence condition there was a sign of a reduction in the increase in success over time during the Social phase compared with the Individual phase (Wald test, β[Phase = social*time] = −0.19, *SE* = 0.09, *Z* = −2.05, *p* = .04). In the Social phase of the Silence condition, the odds of success slowly increased over time (Wald test, β[time] = 0.15, *SE* = 0.09, *Z* = 1.75, *p* = .08).

By contrast, we found a sharper increase in success in Social phase of the Discuss condition (Wald test, β[time] = 0.90, *SE* = 0.18, *Z* = 5.08, $p < .001$), a significant difference from the Individual phase of the Discuss condition (Wald test, β[Phase = individual*time] = −0.50, *SE* = 0.19, *Z* = −2.62, *p* = .009), as well as the Social phase of the Silence condition (Wald test, β[Condition = silence*time] = −0.75, *SE* = 0.20, *Z* = −3.79, $p < .001$).

We found a similar but even clearer pattern of results for the Paul and Linda problem (see Table 2 and Figure 1). During the Individual phase, we found no difference between conditions in either intercept (Wald test, β[silence]- β[discuss] = −0.85, *SE* = 0.67, *Z* = −1.28, *p* = .20) or slope (Wald test, β[Condition = discuss*time] = −0.20, *SE* = 0.19, *Z* = −1.06, *p* = .29). In both conditions the odds of success decreased over time (in the Silence Condition Wald test, β[time] = −0.17, *SE* = 0.13, *Z* = −1.37,

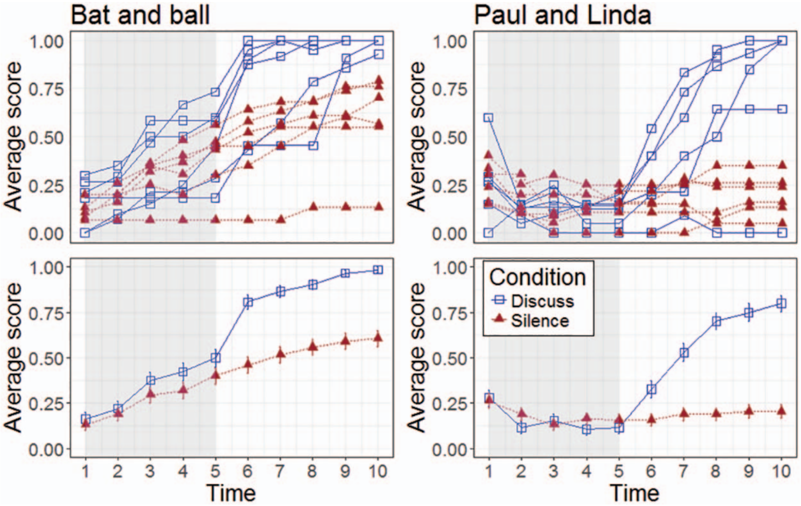

Figure 1. Evolution of the average score during the Individual (shadowed area) and Social phases (clear area, first 5 times steps only) in the Silence (full red triangles [full triangles]) and Discuss (empty blue squares [empty squares]) conditions. Top: average for every group. Bottom: average over all groups. Error bars represent *SE*s. See the online article for the color version of this figure.

p = .17; in the Discuss Condition Wald test, β[time] = −0.37, *SE* = 0.14, *Z* = −2.71, *p* = .007). This decrease likely reflected the fact that participants were pressed to give an answer to a multiple-choice question after 20 s and, therefore, that they were responding almost at chance (chance level was 33%, average success at the first response was 27%).

During the Social phase, there was no sign of improvement over time in the Silence condition (Wald test, β[time] = 0.02, *SE* = 0.13, *Z* = 0.20, *p* = .84). By contrast, we found a sharp increase in success in the Discuss condition (Wald test, β[time] = 0.92, *SE* = 0.15, *Z* = 6.03, *p* < .001), a significant difference from the Individual phase (Wald test, β[Phase = individual*time] = −1.29, *SE* = 0.15, *Z* = −8.67, *p* < .001) and from the Social phase of the Silence condition (Wald test, β[Condition = silence*time] = −0.89, *SE* = 0.20, Z = −4.49, *p* < .001).

Table 2
GLMM Parameters for Paul and Linda's Problem

Analysis of deviance table	χ^2	*df*	*p*-value
Phase	6.51	1.00	.01
Condition	1.91	1.00	.17
Time	.13	1.00	.72
Phase:Condition	2.79	1.00	.09
Phase:Time	43.35	1.00	.00
Condition:Time	2.63	1.00	.11
Phase:Condition:Time	35.16	1.00	.00

Random effects	Variance	*SD*	Corr.
Group	.90	.95	
Group:Time	.06	.25	.93

Fixed effects	Estimate	*SE*	z	*p*-value
Intercept	−2.04	.45	−4.50	.00
Phase (Social)	.28	.32	.88	.38
Condition (Discuss)	−.85	.67	−1.28	.20
Time	−.17	.13	−1.37	.17
Phase (Social):Condition (Discuss)	.88	.49	1.82	.07
Phase (Social):Time	.20	.11	1.86	.06
Condition (Discuss):Time	−.20	.19	−1.06	.29
Phase (Social):Condition (Discuss):Time	1.09	.18	5.93	.00

Note. GLMM = Generalized Linear Mixed Models.

To summarize, our results show that the increase in correct answers is subject to an interaction between phase (Individual vs. Social) and condition (Discuss vs. Silence). There is little evidence for a difference between conditions in the individual phase: the increase in correct answers is nearly the same in the Discuss and Silence conditions. By contrast, in the Social phase, the increase in correct answers is much faster in the Discuss than in the Silence condition.

Remarkably, in one of the discussion group none of the participants had the correct answer at the end of the individual phase. As expected, the group remained stuck on the incorrect answer during the discussion phase, leading to an average score of 0 at the end of the experiment.

Discussion

Discussion among participants enabled the spread of the counterintuitive but correct answer for both problems. As long as some group members had understood the correct answer, they were able to convince their neighbors, who could convince their neighbors in turn until the whole group had accepted the correct answer. The correct answer spread even when it was initially defended by a small minority of participants (2 out of 14 in one group, 3 out of 24 in another), and despite the fact that most of the participants who initially defended the wrong answer did so very confidently (for the participants with an incorrect answer, the modal confidence at the end of the Individual phase was 10—the maximum—

for both problems [Paul and Linda, *Mean* = 8.8; Bat and Ball, *Mean* = 8.0]).

The benefits of discussion are especially striking when compared with the lack of benefits from the mere knowledge of others' answers. Being able to see the other participants' answers yielded no improvement in performance: in neither problem was the rate of increase in correct answers higher in the Social phase than in the Individual phase (in the Silent condition). These results show that discussion can spread the counterintuitive but correct answer in a face-to-face setting. The following experiments were designed to further test the hypothesis that argumentation is necessary and sufficient to spread counterintuitive beliefs.

Study 2: Effect of Repeated Transmission on the Quality of Arguments

Study 2a: Robustness of a Single Argument to Repeated Transmission

While face-to-face argumentation can efficiently spread counterintuitive beliefs, it remains limited in its speed, and potentially in its scope compared with mass media for instance. People acquire and transmit information through formats that can reach large audiences—TV, newspapers—but that do not allow the extensive back and forth of a face-to-face discussion. Can these formats also enable the wide spread of counterintuitive beliefs? For this to be possible, two conditions must be met. The first is that people should be convinced by a single argument instead of a discussion. Previous experiments have shown that a substantial number of participants can be convinced to accept a counterintuitive belief in these conditions (Stanovich & West, 1999; Trouche et al., 2014).

The second condition is that people should be able to convince others in turn, and that the people they convince should be able to convince others, and so forth. This has never been demonstrated so far. To test this, we rely on a variation of the method of transmission chain (e.g., Bangerter, 2000; Bartlett, 1932; Mesoudi & Whiten, 2008) in which a first generation is given an input that it must recall, the recalled input is used as input for another generation, and the process is iterated for several generations. Bartlett used this technique with an argument, but he did not measure the evolution of the persuasiveness of the argument as the transmission events pile up.

In the current experiment, the first generation of participants is given a reasoning problem and asked to provide an answer and an argument defending this answer. Participants of the second generation are asked to solve the same problem, and are provided the answer and argument for the correct answer from a participant from the first generation. They are then given the chance to change their mind, and asked to provide an argument for their final answer. This process is iterated eight times, for a total of eight generations.

Participants. There were 423 participants who were recruited through Amazon Mechanical Turk (161 women, $M_{\text{Age}} = 31.3$, $SD = 10.9$). They were paid $0.5.

Materials. The participants all completed the Paul and Linda problem described above.

Procedure. The first generation was composed of 30 participants who were given the problem, and asked to provide an answer and an argument for their answer.[1] Their justifications were coded according to the following scheme (illustrated by actual answers from participants):

> 0 = Incomplete argument for the good answer (e.g., "Not matter Linda's marital status this would be true").
>
> 1 = Complete argument for the correct answer (e.g., "If Linda is married, she is looking at Patrick, who is not married. If Linda is not married, then Paul (who is married) is looking at her. Thus, in either case, someone who is married is looking at someone who is not married. The answer is Yes.").
>
> 2 = Any incorrect argument for the correct answer (e.g., "I would convince someone that the answer is yes. Paul is already married and Patrick is not married. Linda must also not be married, since she is looking at Patrick. Maybe she desires him. Since that is the case, Paul is looking at someone who is not married.").
>
> 3 = Standard argument for the common wrong answer (e.g., "It doesn't say weather Linda is married, therefore, you can't say for sure.").
>
> 4 = Other arguments for the common wrong answer, and arguments for the other wrong answer (e.g., "We have no idea who people are actually looking at.").[2]

Participants from Generation 2 had to provide an initial answer to the same problem, and were then given an answer and argument presented—truthfully—as coming from a participant in a previous experiment. Four complete arguments for the correct answer (Code 1) were selected at random from Generation 1 and each participant from Generation 2 was randomly assigned to one of these four arguments. Participants then had to give a final answer and a supporting argument. From each group we randomly drew an argument that fulfilled the following two conditions: (a) it was a complete argument for the correct answer (Code 1); (b) it was given by a participant who had initially provided the common wrong answer ("We cannot tell"). Participants from Generation 3 were then randomly assigned to one of these four arguments (see Figure 2 for an example of transmission for one group). The process was iterated until Generation 8 was reached. To sum up, participants at each generation after the first were exposed to a complete argument for the good answer drawn from the answers of participants in the previous generation who had started with the intuitive but wrong answer.

[1] All participants filled in demographic information at the end of the experiment. They were also asked about their confidence in their answer but these results are not discussed further here.

[2] Since the coding appeared straightforward, the authors coded all the justifications. The justifications from the first generation ($N = 162$) were also coded by an external coder, blind to the hypotheses, reaching an intercoder agreement of 0.95.

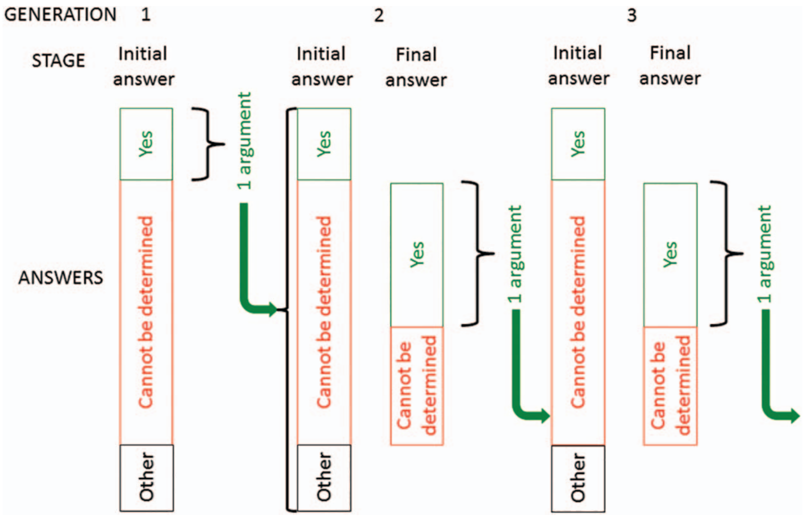

Figure 2. Method for the transmission of arguments in Study 2a. “Yes” refers to people who provide the correct answer (“Yes”), and the correct argument. See the online article for the color version of this figure.

Results

To analyze the results, we rely on two main measures. The first is based on the answers to the problem provided by the participants (i.e., “Yes,” “No,” and “Cannot be determined”). It is the proportion of participants who had started with the intuitive wrong answer (“Cannot be determined”) and changed their mind for the correct answer (“Yes”). This measures whether people have been influenced by the answer and argument provided, without necessarily having understood the argument, and we call it *Influence effectiveness*.

The second measure is based on the justifications provided by the participants. It is the proportion of participants who had started with the intuitive wrong answer (“Cannot be determined”), changed their mind for the correct answer (“Yes”), and provided a correct and complete argument for the correct answer (Code 1). This measures whether people have been able to understand the argument for the correct answer in such a way that they can justify their own correct answer, and we call it *Transmission effectiveness*.

From a cultural evolution perspective these measures play different roles. An argument with a high Influence effectiveness changes the proportion of individuals adopting the correct answer. If the argument changes the mind of enough individuals, it could then have a further impact through conformity. However, the argument itself might not be transmitted, and this might hamper or even preclude further impact if the correct answer needs to be supported by the correct argument to spread. By contrast, an argument with high Transmission effectiveness both influences individuals so they adopt the correct answer, and allows them to repeat the argument adequately and therefore potentially influence more people.

On average, 71% (*SD* = 0.12) of participants initially provided the intuitive but wrong answer. There was no evidence of a change in Influence effectiveness (GLMM Wald test, β[time] = 0.08, *SE* = 0.06, *Z* = 1.30, *p* = .20) or Transmission effectiveness (GLMM Wald test, β[time] = −0.01, *SE* = 0.06, *Z* = −0.23, *p* = .82) over the generations (see Table 3 and Figure 3; details of all the GLMM models are provided in supplementary material). The lack of change in Transmission effectiveness is particularly relevant, because it shows that when participants were convinced by a correct argument, the arguments they put forward to defend the

Table 3
GLMM Parameters for Influence Effectiveness and Transmission Effectiveness

Analysis of deviance table	χ^2	*df*		*p*-value
	Influence effectiveness			
Generation	1.68	1.00		.19
Random effects	Variance	*SD*		Corr.
Intercept	.29	.54		
Generations	.00	.01		−1.00
Fixed effects	Estimate	*SE*	*z*	*p*-value
Intercept	−.16	.38	−.41	.68
Generation	.08	.06	1.30	.20
Analysis of deviance table	χ^2	*df*	*p*-value	
	Transmission effectiveness			
Generation	.05	1.00	.82	
Random effects	Variance	*SD*		Corr.
Intercept	.17	.41		
Generations	.00	.04		−1.00
Fixed effects	Estimate	*SE*	*z*	*p*-value
Intercept	−.31	.34	−.93	.36
Generation	−.01	.06	−.23	.82

Note. GLMM = Generalized Linear Mixed Models.

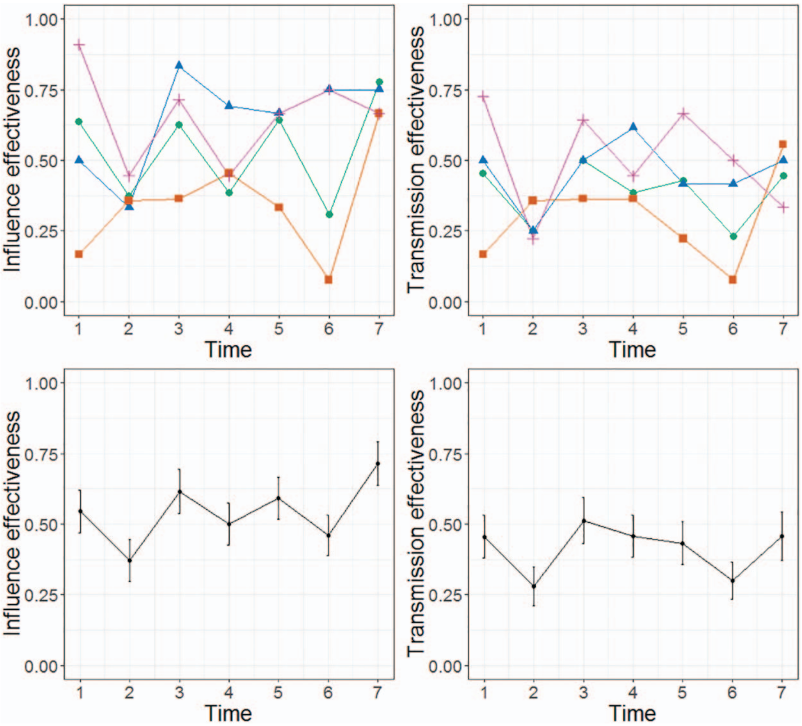

Figure 3. Influence and Transmission effectiveness of the arguments from the four chains of Study 2a. Top: by chain. Bottom: on average. Error bars represent *SE*s. See the online article for the color version of this figure.

correct answer was as convincing as the argument that had convinced them.

This outcome could be obtained through two mechanisms. Participants could memorize the argument they have received with high fidelity, or they could reconstruct the argument on the basis of their newly acquired understanding of the problem. Examination of the arguments provided by the participants suggests that they were not simply memorizing the arguments they had received, since their superficial features were often significantly modified, as in the following example:

> *Argument received:* "If Linda married - She is looking at Patrick who is not married—yes. If Linda not married - Paul is looking at Linda—yes."
>
> *Argument produced:* "There are only two possibilities either Linda is married or she is not married. We know Linda is looking at Patrick and Patrick is not married, so if Linda is married then someone who is married is looking at someone who is not married. What if Lind is not married? Well, Paul is looking at Linda and Paul is married, so if Linda is not married we still have someone who is married, namely Paul, looking at someone who is not married, namely Linda. Either way someone who is married is looking at someone who is not married."

As is apparent in this example, in some cases the arguments produced were significantly more elaborate than the arguments received, strongly suggesting a process of reconstruction. To further demonstrate the importance of reconstruction in argument production, in Study 2b we provided participants with incomplete arguments to test whether participants would then produce equally incomplete arguments or if they would reconstruct the arguments in a more complete form.

Study 2b: Reconstruction of Incomplete Arguments

Study 2a was conservative in that incomplete arguments for the correct answer were not counted as correct answers. However, these arguments can be used to test how much people reconstruct versus memorize the arguments that have convinced them. The prediction is that incomplete arguments will be less convincing, since they are harder to understand, but the participants who are convinced will have had to mentally reconstruct the whole reasoning behind the correct answer. It is then possible that when asked to produce an argument in turn, they produce a complete rather than an incomplete argument.

Participants. There were 295 participants who were recruited through Amazon Mechanical Turk (152 women, $M_{\text{Age}} = 36.0$, $SD = 11.7$). They were paid $0.5.

Method. Participants had to provide an initial answer to the Paul and Linda problem, and were then given an answer and argument presented—truthfully—as coming from a participant in a previous experiment. For half of the participants ($N = 146$), the argument provided was one of three incomplete arguments (Code 0) randomly drawn from the incomplete arguments of Study 2a (e.g., "Not matter Linda's marital status this would be true"). The other participants ($N = 149$) were given a complete argument (Code 1) randomly drawn from the complete arguments of Study 2a. All participants then had to give a final answer and a supporting argument.

Results. There were 80% of participants who initially provided the intuitive but wrong answer. The Influence effectiveness of the incomplete arguments was lower than the influence of the complete arguments (22.2 vs. 49.6%, respectively, $\chi^2(1, N = 236) = 18.0, p < .001$). The Transmission effectiveness was also lower (15.4 vs. 32.8%, respectively, $\chi^2(1, N = 236) = 8.81, p = .003$). However, of the participants who were convinced by the incomplete argument, 69.2% (18 out of 26) produced not only correct but also complete arguments. The proportion of participants producing such arguments was not significantly lower than with a complete argument (66.1% (39 out of 59), $\chi^2(1, N = 85) = 0.001, p = .97$).[3]

To illustrate how participants reconstructed the arguments, here is an example of argument produced by one of the participants who had been convinced by an incomplete argument:

> *Argument received:* Not matter Linda's marital status this would be true.
>
> *Argument produced:* The answer is "Yes someone who is married is looking at someone who is not married." This statement is true regardless of Linda's marital status because if Linda IS married, then her looking at Patrick (who is NOT married) would satisfy the statement; If Linda is NOT married, then Paul (who IS married) looking at Linda would satisfy the statement.

These results show that participants can reconstruct a complete argument from an incomplete one and suggest that reconstruction is also crucial when participants are provided with a complete argument to start with. This process of reconstruction likely explains the remarkable robustness of the arguments across repeated episodes of transmission.

Study 3: Effect of Source Based Biases on the Diffusion of Counterintuitive Beliefs

Studies 1, 2a, and 2b demonstrated the efficacy with which argumentation can spread counterintuitive beliefs. They also suggested that argumentation could potentially overcome other factors such as conformity: in Study 1, the counterintuitive but correct answer defended by a minority of participants spread against the intuitive but incorrect answer defended by the large majority. However, this does not show that social factors such as prestige or conformity could not also spread very counterintuitive beliefs. In the following studies, we test whether conformity and prestige can either contribute to the spread of these beliefs on their own, or assist argumentation in spreading these beliefs.

Study 3a: Effect of Pure Conformity

Participants. There were 156 participants who were recruited through Amazon Mechanical Turk (61 women, $M_{\text{Age}} = 33.0$, $SD = 10.3$). They were paid $0.5.

Method. The methods were similar to those of Study 2b except that, after having given an initial answer, participants were provided with the number of answers of a group of 50 participants described as having previously completed the same problem (instead of being provided with the answer and argument of a previous participant). In the Majority Correct condition, 45 of these participants had answered "Yes" and 5 had answered "We cannot tell." In the Majority Incorrect condition, the numbers were reversed. In two more conditions (Majority Correct with Arguments and Majority Incorrect with Arguments), participants were also provided with one argument supporting each of the two answers, ostensibly given by the previous participants (between-participants 2 × 2 design).

Results. There were 77% of participants who initially provided the intuitive but wrong answer. Compared with Studies 2a and 2b, the most relevant result here lies with the influence of the answers and arguments provided to the participants, irrespective of whether or not they understand the argument (Influence effectiveness).

When only majority information was presented, Influence effectiveness was 0 (i.e., no participant changed their mind) whether the "Yes" (correct) answer was described as being held by a majority or a minority of previous participants (see Table 4). By contrast, introducing the arguments significantly raised Influence effectiveness ($\chi^2(1, N = 120) = 41.4, p < .001$). Participants were more likely to be persuaded when the correct argument was presented as coming from the majority rather than the minority, but this difference was also small and far from significance ($\chi^2(1, N = 61) = 0.87, p = .35$).[4]

Study 3b: Effect of Prestige

Study 3a shows that, in the problems used here, conformity has a very limited effect on the efficiency of arguments and the spread of counterintuitive answers. Study 3b turns to the effects of prestige. If a prestigious individual offers a counterintuitive answer and an argument supporting one might expect participants to accept the answer and the argument, and to transmit the answer and argument to other participants to convince them. The following study, therefore, tested the possibility that prestige—or, more generally, a form of epistemic authority—can lead to the spread of counterintuitive answers and of the arguments supporting them.

Participants. There were 160 participants who were recruited through Amazon Mechanical Turk (59 women, $M_{\text{Age}} = 32.9$, $SD = 10.5$). They were paid $0.5.

Method. The methods were similar to those of Study 2b except that instead of being provided with the answer and argument of a previous participant, participants were explicitly told by the experimenters that they would be given the correct answer to the problem (Pure Prestige condition). In the Prestige and Argument condition, the correct answer (still presented as coming from the experimenters) was accompanied by a correct, complete argument (between-participants design).

[3] To check whether the nonsignificant difference was because of a lack of statistical power, we conducted post hoc power analyses with power $(1 - \beta)$ set at 0.80 and $\alpha = .05$, two-tailed. This showed that our sample size was sufficient to detect an effect size of $h = 0.66$. This suggests that our study could not have detected a small effect. However, the main result is that a strong majority of participants (69.2%) provided a complete argument after having been exposed to an incomplete one, more than the 66.1% who had been convinced by a complete argument.

[4] To check whether this nonsignificant result was because of a lack of statistical power, we conducted post hoc power analysis as in Study 2b. This showed that our sample size was sufficient to detect an effect size of $h = 0.72$. This suggests that our study could not have detected a small effect. However, the main result is that majority information on its own is powerless to change people's mind in this case, not what its effects might be in conjunction with arguments.

Table 4
Effect of Conformity With and Without Argument on Influence Effectiveness

Condition	Influence effectiveness	
	Average value	*SE*
Majority correct	.00	.00
Majority correct with arguments	.62	.09
Majority incorrect	.00	.00
Majority incorrect with arguments	.47	.09

Results. There were 67% of participants who initially provided the intuitive but wrong answer. Because no argument was presented in one of the two conditions, we can only compare Influence effectiveness, and not Transmission effectiveness. Influence effectiveness was significantly higher in the Prestige and Argument condition than in the Pure Prestige condition (Table 5, $\chi^2(1, N = 1{,}007) = 15.2, p < .001$). These results suggest that a strong enough prestige cue can lead people to accept a counterintuitive belief, and that it becomes even more influential in conjunction with a correct argument.

However, few of the participants who accepted the correct answer in the Pure Prestige conditions were able to provide a correct and complete argument for their answer (7 out of 22 participants, 32%). The proportion of complete arguments was smaller in the Pure Prestige condition than in the Prestige and Argument condition (25 out of 40 participants, 63%, $\chi^2(1, N = 22) = 4.19, p = .04$).

Moreover, none of the participants who had accepted the correct answer on the basis of prestige mentioned the source that influenced them to justify their answer (i.e., "The experimenters told us that was the right answer"). This suggests that, in the case at hand, even if prestige and authority might help spread a counterintuitive belief in the coterie of the prestigious source, they would not help spread it any further.

To further support the hypothesis that prestige only has a local effect, we carried out two supplementary studies. Prestige could have an impact beyond those directly influenced by the prestigious individual in two ways. First, even in the absence of argument for the correct answer, many participants accepted the correct answer. Even if very few of these participants provided correct arguments for the correct answer, it could be that their incorrect arguments still help spread the correct answer. The goal of the first supplementary study was simply to confirm that this is not the case. We selected at random three of the incorrect arguments for the correct answer provided by participants who had changed their mind on the basis of prestige only. Out of the 35 participants (from the same population as Studies 2 and 3) who started out with the intuitive but wrong answer, only one ended up with the correct answer and the correct justification.

Even though the participants had not spontaneously provided any argument from authority, they might have done so in a different context. These arguments from authority—of the general form "Authoritative source X said Y"—could then influence other participants, who might repeat these arguments and, therefore, enable the global spread of a counterintuitive belief. In the second supplementary study we checked whether arguments from authority would be efficient to spread the counterintuitive beliefs studied here. We created an argument from authority ("The experimenters told us that was the right answer") and provided it to participants (again from the same population) as being the argument given by another participant in support of the "Yes" answer. Out of the 28 participants (from the same population as Studies 2 and 3) who started with the intuitive but wrong answer, four changed their minds, but none produced a correct argument, and only one repeated an argument from authority. These results thus suggest that the effects of prestige are limited to the participants who are in immediate contact with the prestigious source, and that prestige could not enable the diffusion of the counterintuitive beliefs under study.

General Discussion

Even though intuitive beliefs spread more easily than very counterintuitive beliefs, the latter has still been observed to spread in some cases—scientific and mathematical theories being the most striking example. We hypothesized that argumentation could play a major role in spreading some very counterintuitive beliefs. To test this hypothesis in controlled conditions, we relied on simple reasoning problems that have an intuitive but wrong solution, and a counterintuitive but correct solution. In a series of experiments, we demonstrated the power of argumentation to spread the counterintuitive but correct solution.

Study 1 showed that when participants can discuss the problems together, the correct answer spreads very effectively, even in groups larger than those usually studied. Participants who initially find the correct answer are able to convince the participant they discuss with to accept it. Crucially, the participants who have been convinced can then convince others in turn, until the answer is accepted by the whole group. This is true even when the participants who initially defend the correct answer are a minority, and when they face a confident majority.

Study 2a further demonstrated the robustness of argument transmission. Participants were asked to complete a logical problem, were provided with an argument for the correct answer, could change their mind on this basis, and had to produce an argument for their final answer. The arguments produced were then used as input for another generation of participants, for a total of eight generations. There was no loss of quality in argument effectiveness: the participants who changed their mind produced arguments that were just as convincing as the argument that had convinced them. Study 2b further showed that the robustness of argument transmission was because of the reconstruction of arguments by the participants. Instead of memorizing the argument that had convinced them, the participants reconstructed an argument on the basis of their new understanding of the problem. Participants who managed to understand the problem on the basis of an incomplete argument tended to produce correct and complete arguments.

Table 5
The Effect of Prestige on Influence Effectiveness

Condition	Influence effectiveness	
	Average value	*SE*
Pure prestige	.39	.07
Prestige and argument	.78	.06

Studies 3a and 3b showed that conformity or prestige do not effectively allow the spread of the type of counterintuitive beliefs studied here. When participants were told that a majority of participants had answered in a given way, none changed their mind, unless they were also provided with an argument for the correct answer. Even then, they were not significantly more likely to change their mind than if the argument had been presented without the majority information.

When we told participants, as experimenters, what the correct answer was, approximately half accepted it on the basis of prestige or authority.[5] However, very few participants were then able to produce a convincing argument for the correct answer, and none used arguments that would allow the effects of prestige or authority to be transmitted to another generation of participants. Participants told, explicitly by us, what the correct answer was, but also given an argument for the correct answer, were more likely to then produce correct arguments. However, again none of their arguments referenced the initial source of authority, so that any effect of prestige or authority would be lost after one generation.

Taken together, these results show that argumentation can effectively spread some very counterintuitive beliefs and that, by contrast, conformity and prestige have limited effect on the spread of such beliefs, at least beyond the immediate vicinity of prestigious sources. The complete lack of effect of conformity observed in Study 3a is particularly striking. Note however, that such failures to follow the majority have been observed in other settings (e.g., Efferson, Lalive, Richerson, McElreath, & Lubell, 2008; Efferson et al., 2007) and that this neglect of majority information may be explained by the fact that, at least in some settings, participants tend to give more value to personally acquired information (here their first spontaneous response) compared with social information (see Eriksson & Strimling, 2009; Trouche, Johansson, Hall, & Mercier, 2017). These results suggest that the effects of conformity are strongly modulated by the content of the information being transmitted (see Mercier & Morin, 2017).

In Study 1, the correct answer was defended by a small minority of participants. When conformity and argumentation are pitted against each other in this way, conformity does not significantly hinder the spread of counterintuitive beliefs. This conclusion is reinforced by recent results showing that source based cues—such as the benevolence or the competence of the source—do not stop people from accepting strong enough arguments (Trouche, Shao, & Mercier, 2017; see also Castelain, Bernard, Van der Henst, & Mercier, 2016). This conclusion is also in line with work in the persuasion and attitude change literature showing that when people care about a given belief, people pay more attention to argument quality than to more superficial cues (such as conformity), even if that means accepting unpalatable conclusions (e.g., Petty & Cacioppo, 1979; for review, see Petty & Wegener, 1998).

More generally, along with the results mentioned previously, Study 1 belies the idea that conformity exerts an all-powerful influence (as psychology textbooks sometimes portray it, see Griggs, 2015), showing instead that minorities can manage to spread their beliefs, even if they are counterintuitive (for another demonstration of minority influence, see, e.g., Moscovici & Nemeth, 1974).

To the best of our knowledge, the only past experiment to have investigated the behavior of arguments in transmission chains was that of Bartlett (1932). As mentioned above, in this experiment the arguments were nearly entirely lost after a few generations. This might be explained by the fact that the arguments were relatively long, that they had no particular relevance for the participants, and that the participants were asked to memorize them, not to use them in an attempt to convince someone else. By contrast, in our experiments the arguments where short, were relevant to the participants, and the participants had to produce new arguments to convince someone else. From a methodological point of view, this last point in particular bears emphasizing. Contrary to most prior studies of transmission chains in humans, what was tested here (in particular in Study 2a) was not whether a given piece of information was faithfully reproduced from one generation to the next, but whether this piece of information has the same effect on others from one generation to the next. This focuses the study on the most relevant traits of the piece of information in a given context—here, how convincing of an argument it is.

To better understand the evolutionary dynamic that follows from the present experiments, we constructed simple models based on the results of Studies 2 and 3. These models are useful to extrapolate and generalize from the data obtained in the experiments and to represent situations that fall outside the experimental context. The models represent a population of individuals (fixed at 1,000 individuals) that attempt to solve a counterintuitive problem on their own, and are then exposed to one or more arguments randomly coming from participants from the previous generation. At each generation, the entire pool of individuals changes, so that each generation is composed of 1,000 new individuals. The rate of correct arguments the individuals find on their own (the equivalent of the individual phase of the experiments) is fixed at 10% to approximate the low rate of correct arguments typical of counterintuitive problems.

We study the effects of two variables. The first variable is the Transmission effectiveness of the arguments: the probability that when exposed to a correct argument an initially wrong participant accepts the correct answer and produces a correct argument. The second variable is the number of arguments (N) an individual is exposed to. We assume that if there is at least one correct argument among the arguments participants receive, they are convinced and able to transmit the argument with a certain probability (i.e., Transmission effectiveness). We also assume that individuals exposed to wrong arguments are never convinced to abandon the correct argument (see Laughlin, 2011; Trouche et al., 2014). The main outcome of interest is the proportion of individuals with the correct argument at equilibrium (i.e., when this proportion is stable).

Figure 4 summarizes the findings. When individuals are exposed to a single argument ($N = 1$), only very high Transmission effectiveness (Transmission effectiveness $\geq 90\%$) generates equilibria above 50% (Figure 4B). To understand why even with high Transmission effectiveness the proportion of correct argument remains low, imagine that the simulation starts with an initial population in which everyone has the correct answer. Ten percent of the following generation will find the correct answer on their own. Of the remaining individuals, 810 will be convinced by the correct argument they are exposed to (assuming Transmission

[5] Note that in Experiment 3b the position of authority of the experimenter is in fact much stronger than one of mere "prestige" (sensu Henrich & Gil-White, 2001), so that the case against the role of prestige is even stronger than the one presented here.

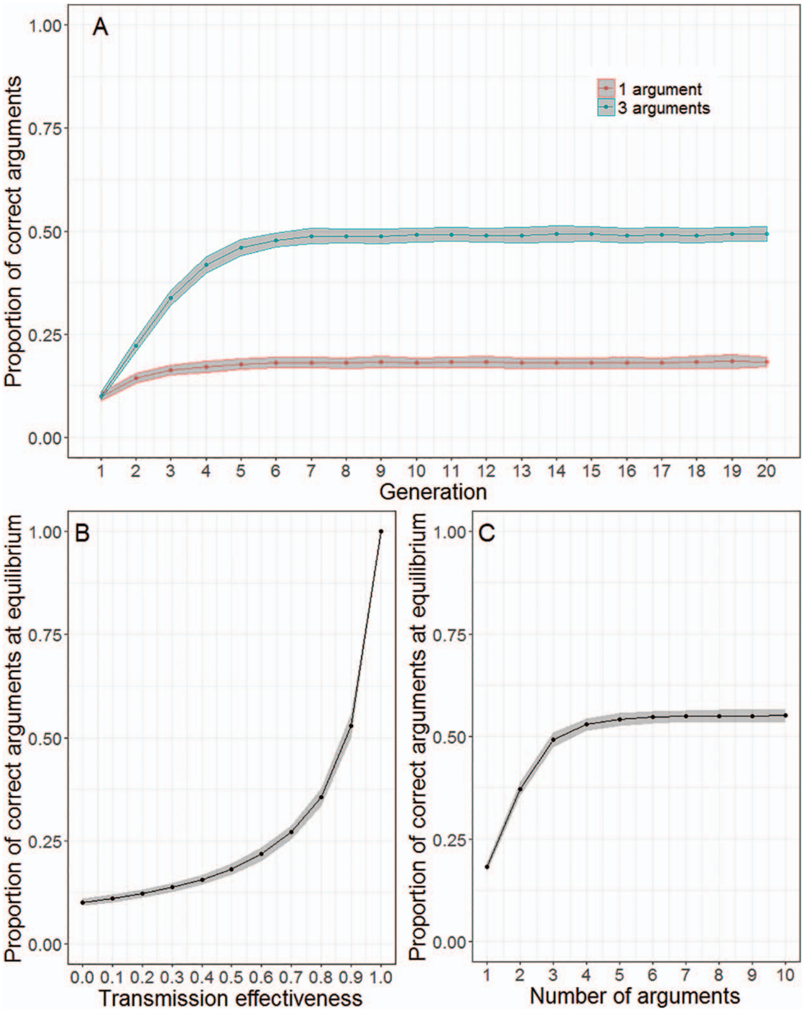

Figure 4. Effect of Transmission effectiveness and the number of arguments (*N*) on the spread of counterintuitive beliefs (proportion of participants giving a correct argument). (A) Increasing the number of arguments participants are exposed to greatly increases the spread of correct arguments (here Transmission effectiveness is fixed at 50%). (B) When the number of arguments is fixed at one ($N = 1$), Transmission effectiveness must be very high to spread the correct arguments. (C) The number of models (*N*) has an important impact on the diffusion of counterintuitive beliefs when transmission effectiveness is limited (Transmission effectiveness = 50%). Dots represent mean values of 100 simulations, the shading the *SD* around the mean. Other parameters of the simulation are the number of individuals (1,000) and the probability to find a correct argument during the initial phase (10%). See the online article for the color version of this figure.

effectiveness = 90%) and the remaining 90 individuals will be left unconvinced and remain with an incorrect argument. Of course, this generation will now provide only 910 correct arguments for the next generation (instead of the 1,000 they received); the proportion of correct arguments, therefore, decreases in the following generations until it reaches equilibrium. This means that for the correct answer to spread significantly along chains of single participants, transmission must be very efficient.

By contrast, when individuals are exposed to three or more arguments an equilibrium above 50% is reached with a relatively modest Transmission effectiveness of 50% (Figure 4A). This is the case simply because the probability that individuals receive one good argument is much higher with repeated sampling of the previous population.[6] Note that when this probability is maximal, that is, when every participant receives all the arguments from the previous generation, the probability that they accept the correct argument is still limited by the transmission effectiveness of the argument (this explains the plateau after four arguments in Figure 4C). This illustrates the importance of redundancy in cultural transmission (see, Acerbi & Tennie, 2016; Enquist, Strimling,

[6] Given the large population size and the limited number of arguments considered, this probability is well approximated by $1\text{-}(1\text{-}p)^n$ where p is the proportion of correct arguments and *n* is the number of arguments drawn. Because p is at least 10%, this probability very quickly converges toward 1 as *n* increases.

Eriksson, Laland, & Sjostrand, 2010; Eriksson & Coultas, 2012; Morin, 2015; Muthukrishna, Shulman, Vasilescu, & Henrich, 2014).

What are the cognitive mechanisms that enable argumentation to spread counterintuitive beliefs? One possibility is that argumentation makes counterintuitive beliefs temporarily intuitive. For instance, with the problems used here participants exposed to the correct argument for the correct answer often immediately and intuitively grasp why the argument is correct and, thus, why the answer it supports is correct (see, e.g., Mercier & Sperber, in press). That is, when participants are confronted with the argument, the correct answer becomes the conclusion of a succession of intuitive inferential steps. Note that this does not mean that the correct answer remains intuitive. On the contrary, the wrong answer often keeps exerting an intuitive pull (see, e.g., Sloman, 1996), and it is likely that participants have to reconstruct the reason for why the correct answer is correct every time they want to convince themselves or someone else of its correctness.

A limitation of our experimental approach is that the counterintuitive answer to the problems we used here might not be as counterintuitive, or not counterintuitive in the same sense, as more culturally significant very counterintuitive beliefs. For instance, it is possible that non-Euclidian geometry violates core ontological intuitions, which is not the case for the correct answer to the Bat and Ball. The study of minimally counterintuitive beliefs faces similar problems in terms of defining the exact way in which these beliefs are counterintuitive (e.g., Atran & Norenzayan, 2004; Purzycki & Willard, 2015). Still, even if finer grained categorizations will eventually be necessary, the category of minimally counterintuitive beliefs has proven to be a very valuable tool. Similarly, finer grained distinctions will have to be made within very counterintuitive beliefs (the rich developmental literature on conceptual change will then prove very helpful, see, e.g., Carey, 1985, 2009; Vosniadou & Brewer, 1992). What the present experiments offer is a first step with material that is more amenable to experimentation than more culturally significant very counterintuitive beliefs—such as non-Euclidian geometry.

Participants who provide the wrong answer to the Bat and Ball, or to the Paul and Linda problem tend to be extremely confident in their wrong answers, making the correct answer quite counterintuitive to them (De Neys, Rossi, & Houdé, 2013; Mata et al., 2013; Trouche et al., 2014). Still, it could be argued that our participants have relatively little commitment to their intuitive but wrong beliefs, which might help explain why argumentation so efficiently spreads the correct but counterintuitive answer. Indeed, it has been suggested that scientists' commitments to their beliefs could be so strong as to make them deeply reluctant to endorse revolutionary theories (e.g., Kuhn, 1962).

Yet, despite these potential difficulties, well-supported scientific theories spread very quickly, even when they are revolutionary. This is particularly true in mathematics. For instance, Gödel's incompleteness theorem was promptly accepted by the mathematical community, even though it disproved beliefs on which eminent members of the community—such as Hilbert or Russell—had wagered their careers (Mancosu, 1999). In the sciences, revolutionary theories also spread very quickly, sometimes only taking a few years to go from eccentric hypotheses to textbook examples (e.g., Oreskes, 1988). Indeed, it has been argued that revolutionary scientific theories spread in the relevant community about as quickly as is warranted by the evidence garnered in their support (Kitcher, 1993; Wootton, 2015). Crucially, as noted above, this is true even if the defenders of the new theories have no special status and are, by definition, a small minority. It is thus possible that our simple experiments capture an important dimension of the spread of some culturally significant very counterintuitive beliefs outside the laboratory. Moreover, our experiments also provide support for the role of argumentation in helping children acquire counterintuitive concepts, as already suggested by the literature on collaborative learning (e.g., Slavin, 1995; on the importance of argumentation, see Henderson, MacPherson, Osborne, & Wild, 2015; Johnson & Johnson, 2009; Mercier, Boudry, Paglieri, & Trouche, in press; Nussbaum, 2008).

References

Acerbi, A., & Tennie, C. (2016). The role of redundant information in cultural transmission and cultural stabilization. *Journal of Comparative Psychology, 130,* 62–70. http://dx.doi.org/10.1037/a0040094

Atran, S., & Norenzayan, A. (2004). Religion's evolutionary landscape: Counterintuition, commitment, compassion, communion. *Behavioral and Brain Sciences, 27,* 713–730. http://dx.doi.org/10.1017/S0140525X04000172

Bangerter, A. (2000). Transformation between scientific and social representations of conception: The method of serial reproduction. *British Journal of Social Psychology, 39,* 521–535. http://dx.doi.org/10.1348/014466600164615

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences, 4,* 29–34. http://dx.doi.org/10.1016/S1364-6613(99)01419-9

Barrett, J. L., & Keil, F. C. (1996). Conceptualizing a nonnatural entity: Anthropomorphism in God concepts. *Cognitive Psychology, 31,* 219–247. http://dx.doi.org/10.1006/cogp.1996.0017

Bartlett, S. F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. Chicago, IL: Chicago University Press.

Boyer, P. (2001). *Religion explained*. London: Heinemann.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (2009). *The origin of concepts*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195367638.001.0001

Castelain, T., Bernard, S., Van der Henst, J.-B., & Mercier, H. (2016). The influence of power and reason on young Maya children's endorsement of testimony. *Developmental Science, 19,* 957–966. http://dx.doi.org/10.1111/desc.12336

Claidière, N., Smith, K., Kirby, S., & Fagot, J. (2014). Cultural evolution of systematically structured behaviour in a non-human primate. *Proceedings of the Royal Society B: Biological Sciences, 281,* 20141541.

Claidière, N., & Sperber, D. (2010). Imitation explains the propagation, not the stability of animal culture. *Proceedings of the Royal Society B: Biological Sciences, 277,* 651–659.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review, 20,* 269–273. http://dx.doi.org/10.3758/s13423-013-0384-5

Efferson, C., Lalive, R., Richerson, P. J., McElreath, R., & Lubell, M. (2008). Conformists and mavericks: The empirics of frequency-dependent cultural transmission. *Evolution and Human Behavior, 29,* 56–64. http://dx.doi.org/10.1016/j.evolhumbehav.2007.08.003

Efferson, C., Richerson, P. J., McElreath, R., Lubell, M., Edsten, E., Waring, T. M., . . . Baum, W. (2007). Learning, productivity, and noise: An experimental study of cultural transmission on the Bolivian Altiplano. *Evolution and Human Behavior, 28,* 11–17. http://dx.doi.org/10.1016/j.evolhumbehav.2006.05.005

Enquist, M., Strimling, P., Eriksson, K., Laland, K., & Sjostrand, J. (2010). One cultural parent makes no culture. *Animal Behaviour, 79,* 1353–1362. http://dx.doi.org/10.1016/j.anbehav.2010.03.009

Eriksson, K., & Coultas, J. C. (2012). The advantage of multiple cultural parents in the cultural transmission of stories. *Evolution and Human Behavior, 33,* 251–259. http://dx.doi.org/10.1016/j.evolhumbehav.2011.10.002

Eriksson, K., & Strimling, P. (2009). Biases for acquiring information individually rather than socially. *Journal of Evolutionary Psychology, 7,* 309–329. http://dx.doi.org/10.1556/JEP.7.2009.4.4

Fehér, O., Wang, H., Saar, S., Mitra, P. P., & Tchernichovski, O. (2009). De novo establishment of wild-type song culture in the zebra finch. *Nature, 459,* 564–568. http://dx.doi.org/10.1038/nature07994

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

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Review. Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 363,* 3503–3514. http://dx.doi.org/10.1098/rstb.2008.0146

Griggs, R. A. (2015). The disappearance of independence in textbook coverage of Asch's social pressure experiments. *Teaching of Psychology, 42,* 137–142. http://dx.doi.org/10.1177/0098628315569939

Henderson, J. B., MacPherson, A., Osborne, J., & Wild, A. (2015). Beyond Construction: Five arguments for the role and value of critique in learning science. *International Journal of Science Education, 37,* 1668–1697. http://dx.doi.org/10.1080/09500693.2015.1043598

Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton, NJ: Princeton University Press.

Henrich, J., & Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and Human Behavior, 19,* 215–241. http://dx.doi.org/10.1016/S1090-5138(98)00018-X

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior, 22,* 165–196. http://dx.doi.org/10.1016/S1090-5138(00)00071-4

Johnson, D. W., & Johnson, R. T. (2009). Energizing learning: The instructional power of conflict. *Educational Researcher, 38,* 37–51. http://dx.doi.org/10.3102/0013189X08330540

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review, 14,* 288–294. http://dx.doi.org/10.3758/BF03194066

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 10681–10686. http://dx.doi.org/10.1073/pnas.0707835105

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 5241–5245. http://dx.doi.org/10.1073/pnas.0608222104

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. New York, NY: Oxford University Press.

Kuhn, T. (1962). *The structure of scientific revolutions* (50th anniversary edition). Chicago, IL: Chicago University Press.

Laland, K. N., & Williams, K. (1997). Shoaling generates social learning of foraging information in guppies. *Animal Behaviour, 53,* 1161–1169. http://dx.doi.org/10.1006/anbe.1996.0318

Laughlin, P. R. (2011). *Group problem solving*. Princeton, NJ: Princeton University Press. http://dx.doi.org/10.1515/9781400836673

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology, 22,* 177–189. http://dx.doi.org/10.1016/0022-1031(86)90022-3

Mancosu, P. (1999). Between Vienna and Berlin: The immediate reception of Godel's incompleteness theorems. *History and Philosophy of Logic, 20,* 33–45. http://dx.doi.org/10.1080/014453499298174

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology, 105,* 353–373. http://dx.doi.org/10.1037/a0033640

Maxwell, R. S. (1936). Remembering in different social groups. *British Journal of Psychology General Section, 27,* 30–40. http://dx.doi.org/10.1111/j.2044-8295.1936.tb00814.x

Mercier, H. (in press). How gullible are we? A review of the evidence from psychology and social science. *Review of General Psychology*.

Mercier, H., Boudry, M., Paglieri, F., & Trouche, E. (in press). Natural born arguers: Teaching how to make the best of our reasoning abilities. *Educational Psychologist*.

Mercier, H., & Morin, O. (2017). Informational conformity: How good are we at aggregating convergent opinions?. Manuscript submitted for publication.

Mercier, H., & Sperber, D. (in press). *The enigma of reason*. Cambridge, MA: Harvard University Press.

Mesoudi, A., & Whiten, A. (2004). The hierarchical transformation of event knowledge in human cultural transmission. *Journal of Cognition and Culture, 4,* 1–24. http://dx.doi.org/10.1163/156853704323074732

Mesoudi, A., & Whiten, A. (2008). Review. The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 363,* 3489–3501. http://dx.doi.org/10.1098/rstb.2008.0129

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to tango: Effects of collaboration and disagreement on dyadic judgment. *Personality and Social Psychology Bulletin, 37,* 1325–1338. http://dx.doi.org/10.1177/0146167211410436

Miton, H., Claidière, N., & Mercier, H. (2015). Universal cognitive mechanisms explain the cultural success of bloodletting. *Evolution and Human Behavior, 36,* 303–312. http://dx.doi.org/10.1016/j.evolhumbehav.2015.01.003

Morin, O. (2013). How portraits turned their eyes upon us: Visual preferences and demographic change in cultural evolution. *Evolution and Human Behavior, 34,* 222–229. http://dx.doi.org/10.1016/j.evolhumbehav.2013.01.004

Morin, O. (2015). *How traditions live and die*. New York, NY: Oxford University Press.

Moscovici, S., & Nemeth, C. (1974). *Social influence: II. Minority influence*. Retrieved from http://psycnet.apa.org/PsycINFO/1974-32232-004

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

Muthukrishna, M., Shulman, B. W., Vasilescu, V., & Henrich, J. (2014). Sociality influences cultural complexity. *Proceedings of the Royal Society of London B: Biological Sciences, 281,* 20132511.

Northway, M. L. (1936). The influence of age and social group on children's remembering. *British Journal of Psychology General Section, 27,* 11–29. http://dx.doi.org/10.1111/j.2044-8295.1936.tb00813.x

Nussbaum, E. M. (2008). Collaborative discourse, argumentation, and learning: Preface and literature review. *Contemporary Educational Psychology, 33,* 15.

Oreskes, N. (1988). The rejection of continental drift. *Historical Studies in the Physical and Biological Sciences, 18,* 311–348. http://dx.doi.org/10.2307/27757605

Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology, 37,* 1915–1926. http://dx.doi.org/10.1037/0022-3514.37.10.1915

Petty, R. E., & Wegener, D. T. (1998). Attitude change: Multiple roles for persuasion variables. In D. T. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 323–390). Boston, MA: McGraw-Hill.

Purzycki, B. G., & Willard, A. K. (2015). MCI theory: A critical discussion. *Religion, Brain & Behavior, 6,* 207–248.

Rahwan, I., Krasnoshtan, D., Shariff, A., & Bonnefon, J.-F. (2014). Analytical reasoning task reveals limits of social learning in networks. *Journal of the Royal Society, Interface, 11,* 20131211. http://dx.doi.org/10.1098/rsif.2013.1211

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition, 111,* 317–328. http://dx.doi.org/10.1016/j.cognition.2009.02.012

Richerson, P. J., & Boyd, R. (2005). *Not by genes alone*. Chicago, IL: University of Chicago Press.

Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting, 12,* 73–89. http://dx.doi.org/10.1016/0169-2070(95)00658-3

Scott-Phillips, T. (in press). A simple (experimental) demonstration that cultural evolution is not replicative, but reconstructive-and an explanation of why this difference matters. *Journal of Cognition and Culture*. Retrieved from http://dro.dur.ac.uk/16582/

Slavin, R. E. (1995). *Cooperative learning: Theory, research, and practice* (Vol. 2nd). London: Allyn & Bacon.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22. http://dx.doi.org/10.1037/0033-2909.119.1.3

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. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1080/07370008.1993.9649028

Trouche, E., Johansson, P., Hall, L., & Mercier, H. (2017). Vigilant conservatism in taking communicated information into account. Manuscript submitted for publication.

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General, 143,* 1958–1971. http://dx.doi.org/10.1037/a0037099

Trouche, E., Shao, J., & Mercier, H. (2017). Objective evaluation of demonstrative arguments. Manuscript submitted for publication.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24,* 535–585.

Wootton, D. (2015). *The invention of science: A new history of the scientific revolution*. London: Harper.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-87458-6

Received April 29, 2016
Revision received April 8, 2017
Accepted April 11, 2017 ■