Adieu Bias: Debiasing Intuitions Among French Speakers

Nina Franiatte, Esther Boissin, Alexandra Delmas, Wim De Neys

Preprint – Psychologica Belgica, 2024 - https://doi.org/10.5334/pb.1260

Adieu Bias: Debiasing Intuitions Among French Speakers

Nina Franiatte*[1,2], Esther Boissin[1], Alexandra Delmas[2], Wim De Neys[1]

[1]Université Paris Cité, LaPsyDÉ, CNRS, 46 rue Saint Jacques, 75005 Paris, France

[2]Research and Development Team, Onepoint, 2 rue Marc Sangnier, 33110 Bègles, France

Abstract

Recent debiasing studies have shown that a short, plain-English explanation of the correct solution strategy can improve reasoning performance. However, these studies have predominantly focused on English-speaking populations, who were tested with problem contents designed for an English-speaking test environment. Here we explore whether the key findings of previous debiasing studies can be extended to native French speakers living in continental Europe (France). We ran a training session with a battery of three reasoning tasks (i.e., base-rate neglect, conjunction fallacy, and bat-and-ball) on 147 native French speakers. We used a two-response paradigm in which participants first gave an initial intuitive response, under time pressure and cognitive load, and then gave a final response after deliberation. Results showed a clear training effect, as early as the initial (intuitive) stage. Immediately after training, most participants solved the problems correctly, without the need for a deliberation process. The findings confirm that the intuitive debiasing training effect extends to native French speakers.

Keywords: Reasoning · Heuristics and biases · Debiasing · Intuition · French

Introduction

Although humans have unique capacities to reason, they are often prone to cognitive biases. Most of the time, people tend to over-rely on fast, intuitive impressions rather than on more demanding, deliberative reasoning when making decisions (Evans, 2003, 2008). This intuitive or so-called 'heuristic' thinking can be useful in many contexts because it is fast, effortless, and often provides valid problem solutions. However, it can also conflict with the most elementary logical or probabilistic principles (e.g., Kahneman, 2011).

For instance, imagine you are analysing the results of a survey in which 1000 people took part. Of the 1000 people, 995 are Americans, and the other 5 are French. You know that one person was drawn randomly from all participants. Next, you are informed that this person loves wine, often goes on strike and has a full month of paid vacation. What do you think is most likely now: Is this person American or French? For many of us, the first response that spontaneously springs to mind is 'a French'. This response is based on stored stereotypical associations cued by the description (e.g., 'French are often perceived as loving wine and going on strike'). If your only piece of information were the description, this answer would probably be correct, as it is likely that there are more French than Americans who love wine and often go on strike. However, if you consider the extreme base-rate information available (995 Americans vs. 5 French), opting for the 'American' option becomes a more compelling choice. Yet untrained people typically neglect the base-rate principle and opt for the intuitive response that is cued by their stereotypical prior beliefs (e.g., Kahneman & Tversky, 1973).

The dichotomy between these two types of responses can be explained by the dual-process model. It characterizes human reasoning as an interplay between two types of processes or 'systems': A fast, intuitive one (often called 'System 1') and a slower, deliberative one (often called 'System 2'; Evans & Stanovich, 2013; Kahneman, 2011). Reasoners who successfully solve the problem in line with standard logico-mathematical principles (e.g., select 'an American' in the above example) would correct their initial intuitive response (e.g., 'a French') after engaging in deliberative calculations (Morewedge & Kahneman, 2010). However, reasoners often refrain from engaging in such calculations. Instead, they default to intuitive processes without considering that the correct answer could be different (Evans & Stanovich, 2013; Kahneman & Frederick, 2005). Hence, as the base-rate example illustrates, relying on mere intuitive thinking can sometimes bias our reasoning (Evans, 2003, 2010; Stanovich & West, 2000).

In many domains, biased judgment can have detrimental impacts (e.g., policy, medicine, law, or education). Against this backdrop, reasoning scholars have long been trying to remediate people's biased thinking (e.g., Lilienfeld et al., 2009; Milkman et al., 2009; Nisbett, 1993). Recent successful debiasing studies have shown that a short training intervention can often help people to reason more

accurately. This intervention consists of a plain-English explanation about the correct solution strategy and the typical biased response to a reasoning problem (e.g., Boissin et al., 2021, 2022; Claidière et al., 2017; Franiatte et al., 2024; Hoover & Healy, 2017; Morewedge et al., 2015; Purcell et al., 2020; Trouche et al., 2014). Typically, in these studies, reasoners who received the intervention are able to produce correct responses to structurally similar problems afterwards.

These recent debiasing training results are promising. However, the nature of the training effect remains unclear. A key question is whether the training primarily affects people's intuitive or deliberate thinking. The common assumption is that after training, participants will be more likely to deliberate properly and engage their 'System 2' to correct the intuitively generated heuristic response (e.g., Evans, 2019; Lilienfeld et al., 2009; Milkman et al., 2009). This idea aligns with the 'corrective' dual-process view which posits that the deliberate 'System 2' primarily serves to correct the intuitive 'System 1' (Kahneman, 2011). However, in theory, it is also possible that once reasoners grasp the solution after the problem is explained, they will no longer generate an incorrect intuitive response. Instead, they might apply the correct solution strategy intuitively without the need for a corrective 'System 2' deliberation process.

Recent evidence provided some support for the 'trained intuitor' viewpoint (e.g., Boissin et al., 2021, 2022). These studies used a two-response paradigm (Thompson et al., 2011) to determine whether the explanation affected participants' intuitive and/or deliberate reasoning. In this paradigm, participants are asked to give two consecutive responses to a reasoning problem. First, they have to respond as fast as possible with the first intuitive hunch that comes to mind. Next, they can take all the time they want to reflect on the problem and give a final response. To make sure that the initial answer is generated intuitively, people have to respond under time pressure and, at the same time, perform a secondary memory task that burdens cognitive resources and disrupts the potential involvement of the deliberative 'system' (Bago & De Neys, 2019). Two-response findings indicate that while the majority of reasoners are biased before the training (both at the initial and final response stages), immediately after receiving the explanation, most of them are able to provide correct responses. Critically, their responses are correct as early as the initial, 'intuitive' stage. This suggests that the debiasing approach allows people to intuit correctly rather than to boost their deliberate correction.

Given that the 'trained intuitor' debiasing approach has important applied and theoretical implications, further validation is needed. However, a critical limitation of (debiasing) training studies is their predominant focus on (native) English speakers, who were tested with problem contents designed for an English-speaking test environment (e.g., Boissin et al., 2021, 2022, 2023a; Franiatte et al., 2024; Hoover & Healy, 2017; Morewedge et al., 2015; Purcell et al., 2020). Such a limited scope of

(most) psychological studies has been questioned by various scholars: They pointed out the lack of representation of diverse populations and languages, urging for greater inclusivity in scientific research (e.g., Arnett, 2008; Blasi et al., 2022; Huettig & Ferreira, 2023; Thalmayer et al., 2021). More specifically, critics have put forward the theoretical and practical limitations stemming from the Anglocentric bias: For instance, the overemphasis on features and mechanisms that are specific to English-speaking people over other populations (see Blasi et al., 2022). Numerous studies also pointed to the overlooked structural differences between languages (e.g., Evans & Levinson, 2009). They can have consequences for other aspects of cognition, ostensibly non-linguistic, such as causal cognition (Bender & Beller, 2019) or biased cognition (Smith et al., 2018).

This critical language limitation raises concerns about the generalizability of recent debiasing findings. Arguably, if we want to guarantee the robustness of the debiasing approach and findings, it seems important to broaden the scope of our research to other languages and cultural settings. As a first step, in the present study, we will test whether the keystone results of previous debiasing studies can be extended to native French speakers living in continental Europe (France).

We used the exact same training procedure and problem test battery as in previous debiasing work (see Franiatte et al., 2024). The only difference was that our participants were native French speakers and all our problem content was adapted to French. The test battery consisted of three popular classic reasoning tasks: The base-rate neglect (Kahneman & Tversky, 1973), conjunction fallacy (Tversky & Kahneman, 1983), and bat-and-ball tasks (Frederick, 2005). They were combined in a one-hour training battery. For each task, the training consisted of three different blocks: A pre-intervention, an intervention, and a post-intervention. Participants were randomly assigned to a training or control group. In the intervention block, participants from the training group solved task problems and always received a short debiasing explanation about the rationale behind the task, while participants of the control group simply solved the problems without receiving the explanation. During the pre- and post-intervention blocks, we used the two-response paradigm to determine whether the intervention affected participants' intuitive and/or deliberate reasoning.

Method

Preregistration and data availability

The study design and research questions were preregistered on the AsPredicted website (https://aspredicted.org) and stored on the Open Science Framework. No specific analyses were pre-registered. Raw data, analysis script, and preregistration are also available on the Open Science Framework (https://osf.io/hk8rv/).

Participants

Participants were volunteers and were recruited online through several communication channels in continental Europe (France).[1] They were not paid for their participation. Only native French speakers were allowed to take part in the study.

In total, 147 reasoners participated in the study (71 females, M_{age} = 37.6 years, *SD* = 12.4), 70 participants were randomly assigned to the training group and 77 to the control group. Among them, 12 had secondary school as their highest level of education, and 135 reported a university degree.

We aimed to recruit a minimum of 100 subjects. Our sample size decision was based on Boissin et al.'s (2021) original study, who tested 100 participants. The experiment ran for three months in the summer of 2022 (early June to late August). We decided to include all participants who had completed the study during that time window. This allowed us to detect small-to-medium training effect (*d* = .41) between the pre- and post-intervention blocks with a power of 80%. All reported results and analyses concern the 147 participants who completed the study.

Materials

The test session was composed of three different reasoning tasks (i.e., base-rate neglect, conjunction fallacy, and bat-and-ball tasks). In each session, for each participant, the task order was randomized. Each task contained eight conflict and eight no-conflict problems (see further) and was composed of three blocks presented in the following order: A pre-intervention, a short intervention, and a post-intervention block. In total, each participant had to solve 48 problems. All these problems had been adapted in French before running the experiment (see Procedure). Problems are presented in Supplementary Material Section A. For convenience, we will always illustrate the problem content in the main text with English examples.

Base-rate neglect problems (BR).

Each participant was presented with base-rate problems based on Pennycook et al. (2014) that were already adapted in French by Boissin et al. (2023b). Participants always received a description of the composition of a sample (e.g., 'This study contains writers and construction workers'), a description that was designed to cue a stereotypical association (e.g., 'Person 'W' is

[1] After verification, one participant took the experiment in Québec, Canada. Since this individual identified as a native French speaker and their performance aligned with the overall trend, we decided not to exclude them from the analyses.

strong'), and a base rate information (e.g., 'There are 996 writers and 4 construction workers'). Participants' task was to indicate to which group the person most likely belonged. The task instructions stressed that the person was drawn randomly from the specified sample. The problem presentation format was based on Pennycook et al.'s (2014) rapid-response paradigm. The base rates and descriptive information were presented serially, and the amount of text presented on screen was minimized. As in Pennycook et al. (2014), base rates varied between 995/5, 996/4, and 997/3. We labelled the response that is in line with the base rates as the correct response (see Supplementary Material Section A). Table 1 illustrates the full problem format.

To ensure that possible correct or incorrect responses did not originate from guessing, we also presented no-conflict control problems. In these control problems, the description always triggered a stereotypical trait of a member of the largest group. The heuristic intuition thus cued the correct response. Participants had to select the correct response among the same two answer options as for a corresponding standard conflict version (see Table 1).

We presented four conflict and four no-conflict problems in the pre- and post-intervention blocks. These no-conflict problems should be easy to solve. If participants are paying minimal attention to the task and refrain from random guessing, they should show high accuracy (Bago & De Neys, 2019).

Conjunction fallacy problems (CF).

Each participant was also presented with conjunction fallacy problems. Our item material was based on a new pilot study in which we adapted and pretested conjunction fallacy problems in French (see *Pilot rating study* in Procedure). We used the conjunction task format introduced by Andersson et al. (2020): All conjunction problems presented a short personality description of a character, consisting of their name (e.g., 'Kadin'), their age (e.g., '32'), their previous studies (e.g., 'astronomy') and their hobby/interest (e.g., 'sci-fi'). Next, the participants were given four response options and were asked to indicate which one was most likely. In the critical conflict problems, one option presented a characteristic that featured an unlikely stereotypical association given the description (e.g., 'a longshoreman'), and one option presented a conjunction of this unlikely and a likely characteristic (e.g., 'a longshoreman and a stargazer'). Two other filler options presented a very unlikely characteristic (e.g., 'an Oscar winner') and a conjunction of two unlikely characteristics (e.g., 'a longshoreman and an equestrian'). Table 1 illustrates the full problem format.

We also presented four conflict and four no-conflict control problems in the pre- and post-intervention blocks. In the no-conflict control problems, we replaced the singular unlikely response

option with the option that featured the likely stereotypical association (e.g., 'a stargazer' in the above example, see Table 1). Reasoners will tend to select the statement that best fits with the stereotypical description (Tversky & Kahneman, 1983). The fit will be higher for the likely than the unlikely characteristic with the conjunctive statement falling in between. Hence, on the no-conflict problems, stereotypical associations will no longer favour the conjunctive over the singular statement and participants are expected to show high accuracies (see De Neys et al., 2011).

The four response options were presented in random order. Note that Andersson et al. (2020) adopted the four options design to minimize the use of simple visual response strategies (e.g., 'always choose the shortest answer'). As in the Andersson et al. study, selection of the filler options was overall low in our study (i.e., 17.4% of options). However, strictly speaking, participants who select the singular very unlikely option (e.g., 'an Oscar winner' in the above example) do not violate the critical conjunction rule. As Boissin et al. (2022) mentioned, given that we are interested in learning effects, selection of the very unlikely option can be considered a correct response. Hence, we considered answers on which the conjunction fallacy is avoided (i.e., unlikely and very unlikely answers) as correct answers. Figure S1 in Supplementary Material Section B gives a detailed overview of the selection frequency of each individual response option.

Bat-and-ball problems (BB).

We also presented problems taken from Raoelison and De Neys (2019). They were modified, French versions of the original bat-and-ball problem (Frederick, 2005), which used quantities instead of prices (e.g., '*In a park there are 140 adults and children in total. There are 100 more adults than children. How many children are there?')*. Participants had to select the correct response among four response choices which were composed of (1) the correct response (i.e., '20 children' in the above example), (2) the intuitively cued 'heuristic' response (i.e., '40 children' in the above example), (3) a foil option which was the sum of correct and heuristic answers (i.e., '60 children'), and (4) a second foil option which was the second greatest common divider (i.e., '10 children'). Mathematically speaking, the correct equation to solve the above bat-and-ball problem is: '100 + 2x = 140'. Instead, people are thought to be intuitively using the '100 + x = 140' equation to determine their response (Kahneman, 2011). The latter equation was used to determine the 'heuristic' answer option, and the former to determine the correct answer option for this problem. The four response choices appeared in random order. Table 1 illustrates the full problem format.

We also presented four conflict and four no-conflict control problems in the pre- and post-intervention blocks. In the no-conflict control problems, the conflict was removed by deleting the critical relational 'more than' statement. The heuristic intuition thus cued the correct response (see

Table 1; De Neys et al., 2013). In this case, the intuitively cued '40 children' answer was correct. Note that, as Boissin et al. (2021), we added three words to the control problem questions to equate the semantic length of the conflict and no-conflict versions. Participants had to select the correct response among the same four answer options as for a corresponding standard conflict version. As in the other tasks, these control problems should be easy to solve: If participants are paying minimal attention to the task and refrain from random guessing, accuracy should be at ceiling (Bago & De Neys, 2019).

	Conflict version	**No-conflict version**
Base-rate neglect	This study contains writers and construction workers. Person 'W' is strong. There are 996 writers and 4 construction workers. Is Person 'W' more likely to be: • A writer • A construction worker	This study contains writers and construction workers. Person 'W' is strong. There are 996 construction workers and 4 writers. Is Person 'W' more likely to be: • A writer • A construction worker
Conjunction Fallacy	Kadin, 32, has previously studied astronomy and likes sci-fi. Is it most probable that the described person is: • A longshoreman • A longshoreman and a stargazer • An Oscar winner • A longshoreman and an equestrian	Kadin, 32, has previously studied astronomy and likes sci-fi. Is it most probable that the described person is: • A stargazer • A longshoreman and a stargazer • An Oscar winner • A longshoreman and an equestrian
Bat-and-ball	In a park, there are 140 adults and children in total. There are 100 more adults than children. How many children are there? • 40 children • 10 children	In a park, there are 140 adults and children in total. There are 100 adults. How many children are there in the park? • 40 children

	• 60 children • 20 children	• 10 children • 60 children • 20 children

Table 1. Examples of conflict and no-conflict problems for the three reasoning tasks used in the battery: Base-rate neglect, Conjunction Fallacy, and Bat-and-ball. For convenience, the problem content is illustrated in English. The corresponding French material can be found in the Supplementary Material Section A.

Counterbalancing.

For every reasoning task, two sets of problems were created in which the conflict status of each problem (see above) was counterbalanced. More specifically, all the conflict problems of the first set appeared in their no-conflict version in the second set, and vice-versa. Half of the participants were presented with the first set of problems, while the other half was presented with the second set. Hence, in each task, the same content was never presented more than once to a participant, and everyone was exposed to the same problems, which minimized the possibility that mere problem differences influence the results. The presentation order of the tasks and the problems within each task was also randomized.

Intervention block.

In the intervention block, participants had to solve three additional conflict problems (i.e., three base-rate or three conjunction fallacy or three bat-and-ball problems depending on the task), without any cognitive or time constraint. In the training group, participants were explained the correct solution after having responded to each problem, whereas in the control group, participants only responded to the problem without receiving any explanation. The explanations were translated from English to French. They were based on the same general principles that were adopted by Boissin et al. (2021, 2022): They were as brief and simple as possible to prevent fatigue or disengagement from the task. Each explanation explicitly stated both the correct response and the typical biased, incorrect response. No personal performance feedback was given to avoid promoting feelings of judgment (Trouche et al., 2014). Finally, to avoid inducing mathematical anxiety, the explanation never mentioned a formal algebraic equation (Hoover & Healy, 2017). The following example illustrates a typical question and explanation for a bat-and-ball problem:

'Question:

A banana and an apple cost $1.40 in total. The banana costs $1.00 more than the apple. How much does the apple cost?

Explanation:

The correct response is 20 cents. Many people are tempted to answer 40 cents, but this is wrong.

If the apple costs 40 cents, the banana would cost $1.40 (as it costs one dollar more than the apple); both together, they would then cost $1.80.

However, the problem said they cost $1.40 together.

The correct answer is that the apple costs 20 cents, the banana $1.20 so together they cost $1.40 ($0.20 + $1.20 = $1.40).'

Two-response format.

We used the two-response paradigm (Thompson et al., 2011) for the presentation of all problems in the pre- and post-intervention blocks. In this paradigm, participants are asked to provide two consecutive responses on every trial: A 'fast' response, directly followed by a second 'slow' response. This method allowed us to capture both an initial 'intuitive' response, and then a final 'deliberate' one. To minimize the possibility that deliberation was involved in producing the initial 'fast' response, participants had to provide their initial answer within a strict time limit while performing a concurrent cognitive load task (e.g., Bago & De Neys, 2017, 2019). The cognitive load task was based on the dot memorization task (Miyake et al., 2001) given that it had been successfully used to burden executive resources during reasoning tasks (e.g., De Neys, 2006; Franssens & De Neys, 2009). Participants had to memorize a complex visual pattern (i.e., a 3 x 3 grid in which 4 dots were placed) that was presented briefly before each reasoning problem. After their initial 'intuitive' response to the problem, participants were shown four different matrixes, and they had to choose the correct pattern (see De Neys, 2006, for more details). They received feedback as to whether they chose the correct or incorrect pattern.

For all base-rate problems, a time limit of 3 seconds was chosen for the initial response, based on previous pre-testing that indicated it amounted to the time needed to read the preambles, move the mouse, and click on a response option. Similarly, the time limit was set to 5 seconds for conjunction fallacy problems and 8 seconds for bat-and-ball problems. For all tasks, previous pretesting established that the time limits imposed a stringent time pressure that forced participants to respond significantly faster than in a traditional unconstrained, one-response test format (Bago & De Neys, 2017, 2019; Boissin et al., 2022). Note that the time limit and cognitive load were only applied during the initial response stage and not during the subsequent final stage in which participants were allowed to deliberate.

Justification.

For every reasoning task, after the last problem of the post-intervention block - which was always a conflict problem - participants were asked to select a rationale for their final response (they could choose between: '*I did the math*' / '*I guessed*' / '*I decided based on intuition or gut feeling*' / '*Other*'). For the '*Math*' and '*Other*' options, they were asked to type-in an explanation for their justification. Previous work (e.g., Bago & De Neys, 2019; Boissin et al., 2021) indicated that correct reasoners typically manage to correctly justify their answers.

As in previous studies, results indicated that, for the three tasks, the majority of correct responses was correctly justified after training (training group: 105 correct justifications out of 170 correct responses, i.e., 61.8%; control group: 64 correct justifications out of 96 correct responses, i.e., 66.7%). The interested reader can find details in Table S1 in Supplementary Material Section C. Note that the justification was untimed and retrospective. It was collected for exploratory purposes and does obviously not allow drawing any conclusions regarding the intuitive or deliberate nature of participants' processing.

Procedure

Pilot rating study.

The material for the base-rate and bat-and-ball task items was already adapted to French and validated in previous pilot studies (e.g., Boissin et al., 2023b; Raoelison et al., 2021). For the conjunction task, we created a pool of 52 potential French items that contained translated and culturally adapted items from Andersson et al. (2020) and newly generated items that respected the same structure. To validate the stereotypical problem content, we ran a pilot rating study with 90 participants (45 females, 2 neutral-gender, M_{age} = 30.2 years, *SD* = 9.8). Participants were asked to rate how well each option matched the described person on a scale from 0 (not at all similar) to 10 (very similar). To select the most appropriate material, after an initial exploration, we picked items for which, in the conflict version, the combination of the unlikely and likely constituent was rated at a minimum of 3.5 and was rated higher than the unlikely constituent. In their no-conflict counterpart, we picked items for which the likely constituent was rated at a minimum of 5 and higher than the combination of the unlikely and likely constituent. In addition, the relative option ranking needed to be maximally respected (e.g., very unlikely < unlikely < likely and unlikely combination < likely). We selected 35 items for which these differences were greatest. Among the ultimately selected items, the average ratings for the different response options were: Very unlikely option (*M* = 0.6, *SD* = 0.7); unlikely option (*M* = 1.6, *SD* = 1.5); unlikely and unlikely option (*M* = 1.3, *SD* = 1.3); unlikely and likely option (*M* = 5.1, *SD* = 1.8); and likely option (*M* = 7.1, *SD* = 1.6). In total, the 35 items were distributed

as follows: 2 counterbalanced sets of 8 items in each pre- and post-intervention blocks and 3 items in the intervention block. The full item set can be found in Supplementary Material Section A.

Main study.

The experiment was conducted online using the Qualtrics platform (https://www.qualtrics.com), either in small groups in the presence of an experimenter or at home. The procedure was similar to Franiatte et al. (2024). First, participants were instructed that the experiment would take around fifty-five minutes and that it demanded their full attention. They were told they would need to solve different types of reasoning tasks for which they would have to provide two consecutive responses. They were specifically instructed that we were interested in their very first, initial answer that comes to mind and that – after providing their initial response – they could reflect on the problem and take as much time as they needed to provide a final answer. At the beginning of each task, to familiarize themselves with the two-response procedure, they solved two unrelated practice reasoning problems. Next, they familiarized themselves with the cognitive load task by solving two load trials and, finally, they solved two problems which included both cognitive load and the two-response procedure.

Figure 1 shows a typical trial, which consisted of, first, presentation of a fixation cross displayed during 2000 ms, followed by the first sentence of the problem displayed for 2000 ms (e.g., *'In a park, there are 140 adults and children in total'* for the bat-and-ball task), and followed by the visual matrix for the cognitive load task for 2000 ms. Then, the full problem was presented, at which point participants had 3000 ms (base-rate neglect), 5000 ms (conjunction fallacy), or 8000 ms (bat-and-ball) to give their initial answer. Note that in this initial 'intuitive' response stage, the background of the screen turned yellow after 2000 ms (base-rate neglect), 3000 ms (conjunction fallacy), or 6000 ms (bat-and-ball) to warn participants that they only had a short amount of time left to answer. If they had not provided an answer before the time limit, they were given a reminder that it was important to provide an answer within the time limit on subsequent trials. Participants were then asked to enter their confidence in the correctness of their answer on a scale from 0% (absolutely not confident) to 100% (absolutely confident). Then, they were presented with four visual matrix options and had to choose the one that they had previously memorized. Finally, the same reasoning problem was presented again, and participants were asked to provide a final 'deliberate' answer (without time limit nor cognitive load) and, once again, to indicate their confidence level. Note that due to a coding error the confidence data was not systematically recorded and was not further analysed (the non-missing data is included in our data file, and an exploratory analysis of the partial data can be found in Supplementary Material Section D).

At the end of the study, participants in the control group were also presented with the explanations about how the base-rate neglect, conjunction fallacy, and bat-and-ball problems could be solved, and all participants were asked to complete a page with demographic questions.

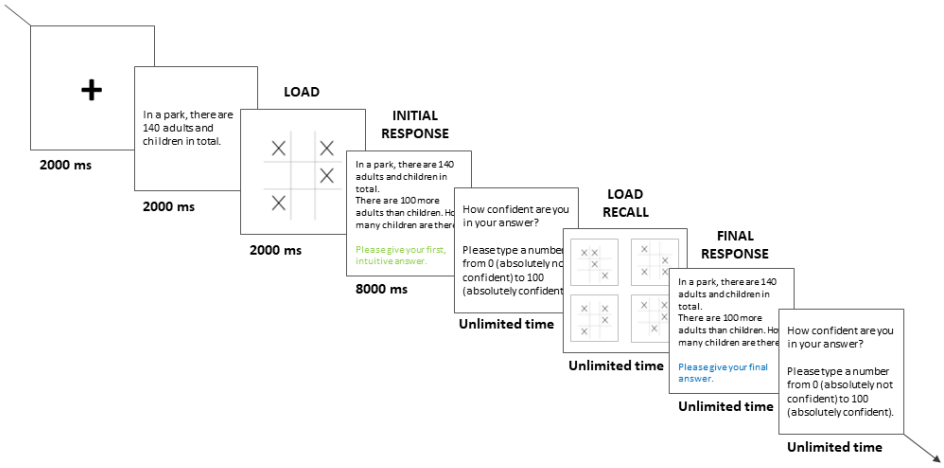


Figure 1. Time course of a typical two-response trial, with a bat-and-ball problem. For convenience, the problem content is illustrated in English.

Trial exclusion

Following our preregistration, we discarded trials in which participants failed to provide their initial answer before the deadline (5.5% of all trials) or failed to pick the correct matrix in the cognitive load task (9.6% of the remaining trials), and we analysed the remaining 90.4% of all trials. On average, each participant contributed 40.9 (*SD* = 5.1) conflict trials out of 48, and 41.0 (*SD* = 4.3) no-conflict trials out of 48.

Note that as part of our procedure, we asked participants whether they were familiar with the original bat-and-ball problem and asked them to solve it (Frederick, 2005). In total, 95 participants out of 147 (64.6%) reported having come across the problem before. Traditionally, these participants are removed from the analyses to eliminate the possibility that their prior knowledge of the correct solution affects the results (e.g., Bago & De Neys, 2019; Boissin et al., 2021). First, we ran all analyses while including these 95 participants, and second, while not including them. None of our conclusions were affected either way, and the tendencies remained the same. Thus, in line with our preregistration, we included these participants in the reported analyses in the main text (see Figure S2 in Supplementary Material Section E for overview analyses with and without these participants).

Composite measure

For simplicity and to maximize power, our analyses focused on the composite conflict accuracy across the three different reasoning tasks (i.e., base-rate neglect, conjunction fallacy, and bat-and-ball). To calculate the composite performance, we averaged for each participant the proportion of correct initial and final responses, separately for each task. Then we averaged across all tasks (separately for initial and final trials). For completeness, we calculated the composite performance also for no-conflict trials (see Table S3 in Supplementary Material Section F).

Statistical analyses

The data were processed and analysed using the R software (R CoreTeam, 2017) and the following packages (in alphabetical order): dplyr (Wickham et al., 2020), ggplot2 (Wickham, 2016), lmerTest (Kuznetsova et al., 2017) and tidyverse (Wickham et al., 2024).

The study uses a between-subject group variable based on the assigned group during the experiment (i.e., training or control). We measured the percentage of correct responses – referred to as the accuracy in the main text - on conflict and no-conflict items, at both initial 'intuitive' and final 'deliberate' response stages, before (pre-intervention) and after (post-intervention) receiving the text training.

Throughout the article, we used mixed-effect regression models with random intercepts and slopes for both participants and stimuli (i.e., items**;** see Brysbaert & Debeer, 2023). The Wald test assessed the statistical significance of the fixed effect of the model.

Results

Conflict trial accuracy

For each task and for each participant, we analysed the average proportion of correct initial and final responses for all the conflict items, in each of the two blocks (pre- and post-intervention). First, before the intervention, participants were mostly biased and showed low initial accuracies (training group: M = 37.8%, SD = 28.2; control group: M = 32.8%, SD = 24.5; see Figure 2). The overall performance of both groups improved following the intervention. However, the accuracy increase was significantly higher in the training group (+45.0 points, M = 82.8%, SD = 24.2) than in the control group (+6.4 points, M = 39.2%, SD = 25.5). Statistical composite analyses revealed that the Block x Group interaction significantly improved the model for the initial responses, $\chi 2$ (1) = 103.2, p < .001.

In the same vein, participants showed lower final accuracies before the intervention (training group: *M* = 51.5%, *SD* = 25.5; control group: *M* = 42.8%, *SD* = 25.2) than after. In the post-intervention block, participants of the training group sharply improved their performance (+32.7 points, *M* = 84.2%, *SD* = 21.7), while those of the control group hardly improved (+1.4 points, *M* = 44.2%, *SD* = 26.0). Similarly, statistical composite analyses revealed that the Block x Group interaction significantly improved the model for the final responses, $\chi 2$ (1) = 88.2, $p < .001$. The interested reader can find details of the main effects in Tables S4 and S5 (Supplementary Material Section G).

For completeness, Figure 2 (bottom panels) shows the data for each individual reasoning task. By and large, similar initial and final response tendencies were observed for each individual task. If anything, as previously found in Franiatte et al. (2024), the training effect tended to be somewhat less pronounced for the base-rate task. However, in this task, participants' pre-intervention performance was also already higher than for the others.

Additionally, we conducted an exploratory analysis to compare our study's reasoning performance with the similar debiasing study—adopting the same tasks and design—conducted by Franiatte et al. (2024) on native English speakers. Overall, we observed a similar training effect in the two samples (see Figure S3 in Supplementary Material Section H). In the training group, initial responses significantly improved by 45% in the current study, and 42% in Franiatte et al.'s (2024) study. Similarly, the final responses significantly improved by 33% in the current study and 44% in Franiatte et al.'s (2024) study. Tendencies for the individual reasoning tasks were also highly similar (see Supplementary Material Section G).

In sum, in our study, we replicated previously established findings in a native French-speaking sample of reasoners. Our results are consistent with the recent debiasing literature (e.g., Boissin et al., 2021, 2022; Franiatte et al., 2024; Purcell et al., 2022) and confirm that a single, short intervention can significantly increase both initial and final response accuracy on classic reasoning tasks.

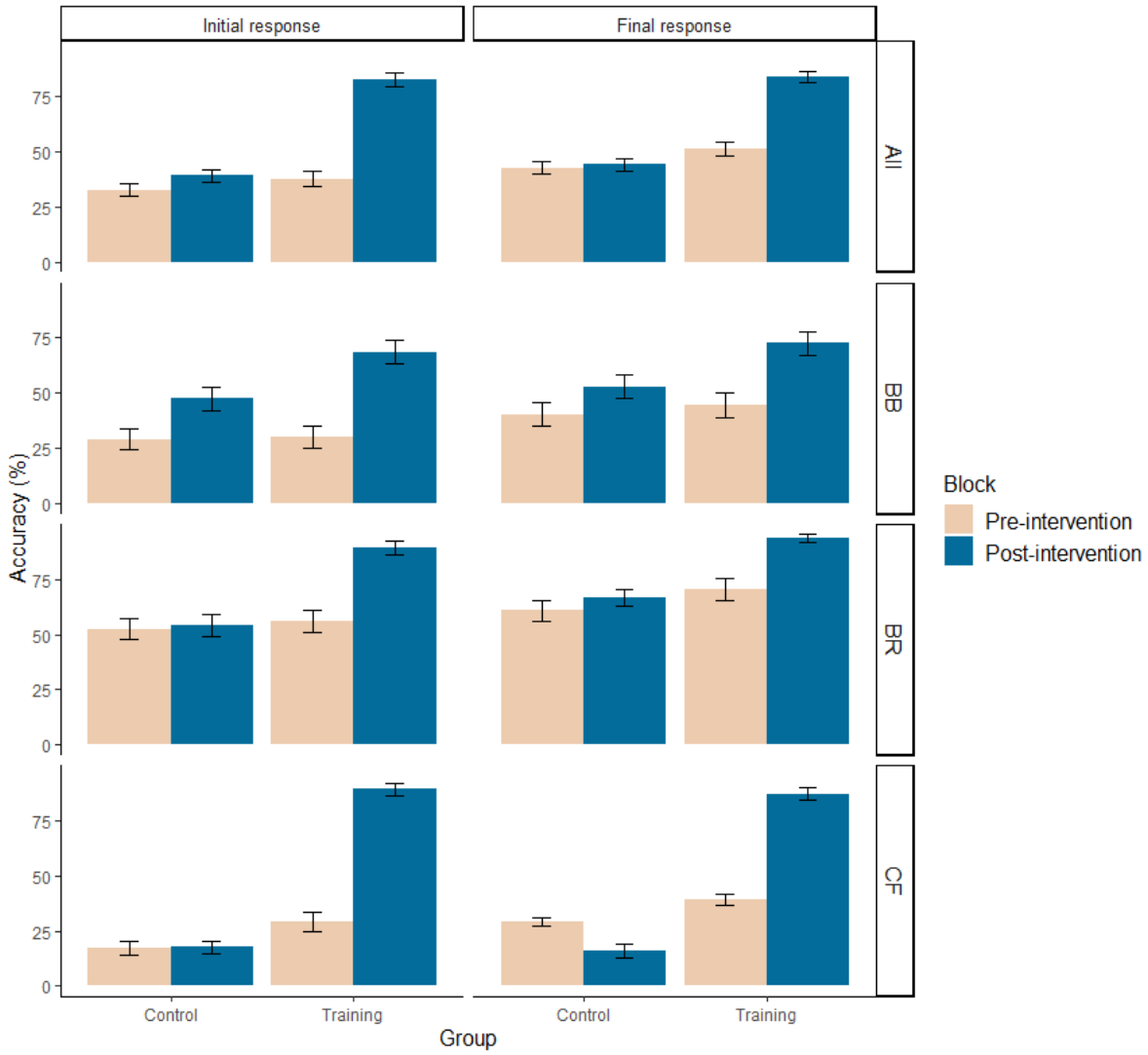


Figure 2. Mean accuracy (%) of correct initial and final responses on conflict problems for control and training groups, before and after the intervention, for each task (BB, BR, CF), and combined (All). Error bars are standard errors. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks, All = the composite mean across the three tasks.

Direction of change

To gain some insight into how people changed (or did not change) their answers after deliberation, we performed a direction of change analysis for the conflict items (Bago & De Neys, 2017). Specifically, each trial is composed of two responses, the initial 'intuitive' one (given under time pressure and cognitive load) and the final 'deliberate' one. Correct responses are labelled '1' and incorrect responses are labelled '0'. Hence, each trial can result in one of four different patterns: '00' pattern, incorrect response at both response stages; '11' pattern, correct response at both response stages; '01' pattern, initial incorrect and final correct responses; and '10' pattern, initial correct and final incorrect responses. Figure 3 plots the direction of change distribution for each block (pre- and post-intervention) and each group (control and training).

Consistent with the overall accuracies presented above, before the intervention, around half of the trials were incorrect at both response stages and produced '00' (biased) patterns (training group: *M* = 43.0%, *SD* = 27.1; control group: *M* = 51.1%, *SD* = 26.1). After the intervention, similar results were observed for participants in the control group, with '00' (biased) patterns remaining stable (0.1 point decrease, *M* = 51.0%, *SD* = 27.07 in the post-intervention block). However, in the training group, the intervention led to a sharp decrease in '00' patterns (31.5 points decrease, *M* = 11.5%, *SD* = 18.8). Notably, the decrease in '00' patterns led to a considerable increase in '11' patterns (41.9 points rise in the training group vs. 6.1 points rise in the control group) rather than in '01' patterns wherein we observed the opposite trend (13.5 points decrease in the training group; 5.9 points decrease in the control group). Eyeballing Figure 3 (bottom panels) indicates that we observed similar tendencies for each of the individual reasoning task. These tendencies were again highly consistent with the original Franiatte et al.'s study with English speakers (see Table S6 in Supplementary Material Section I for full details).

In sum, these results confirm that the training improved reasoning performance, as early as the initial 'intuitive' stage. In this study, we replicated the sound intuiting effect found in previous debiasing studies (e.g., Boissin et al., 2021, 2022). In other words, after the training intervention, reasoners were able to intuit the correct solution strategy and typically no longer required to correct an initial 'erroneous' response through deliberation.

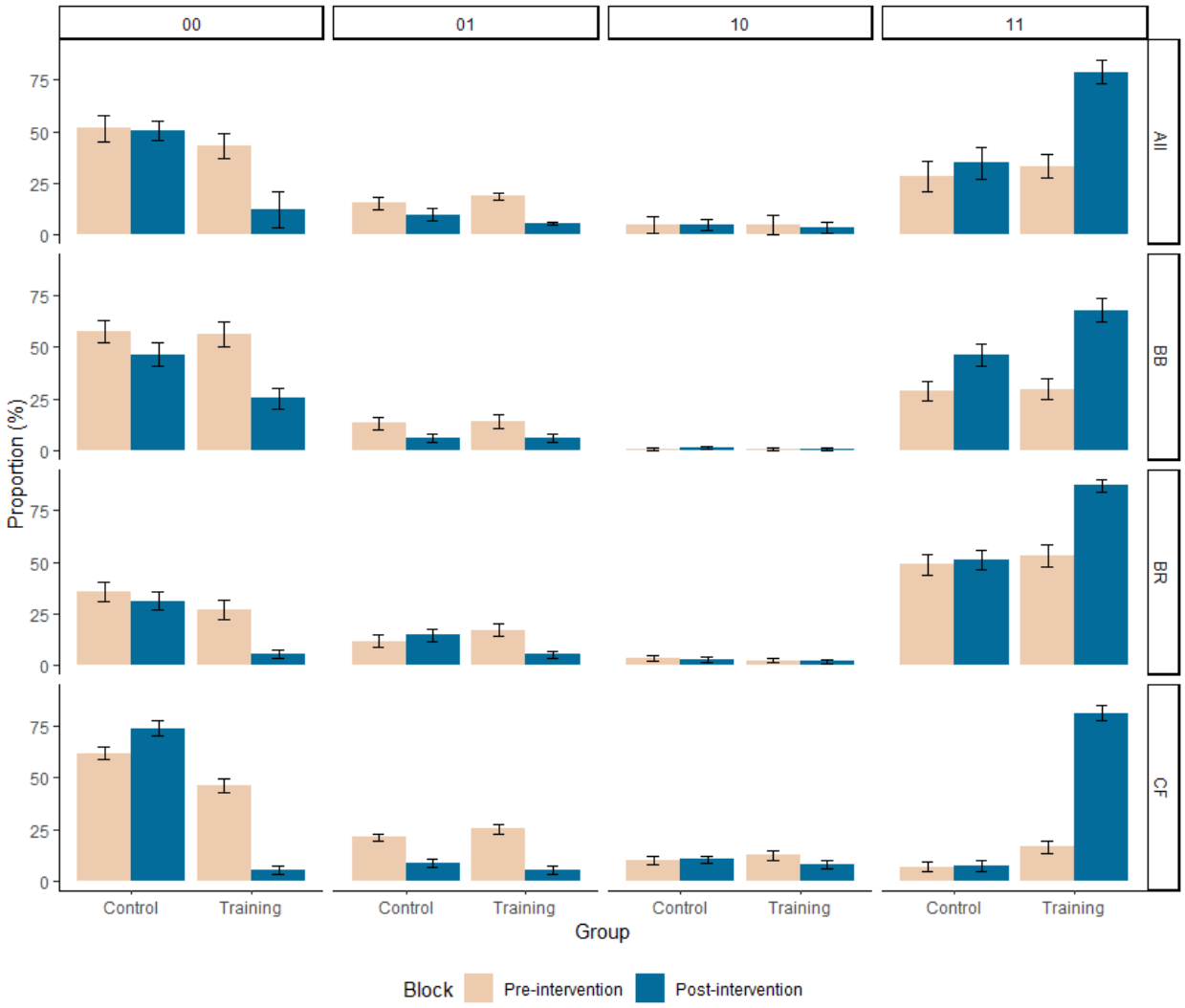


Figure 3. Proportion (%) of each direction of change (i.e., '00' pattern, '01' pattern, '10' pattern, and '11' pattern; 0 = incorrect response, 1 = correct response, first digit = initial response, second digit = final response) on conflict problems for control and training groups, before and after the intervention, for each task (BB, BR, CF), and combined (All). Error bars are standard errors. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks, All = the composite mean across the three tasks.

Individual level direction of change

To gain some deeper insight into how a given reasoner changed (or did not change) their response, we also performed an individual level accuracy analysis on the conflict trials (Raoelison & De Neys, 2019). For each of the 147 participants, on each conflict trial, from start to end of the experiment, we focus on their dominant direction of change and classified it using the categories introduced by Boissin et al. (2021, 2022).

First, 'biased responders' did not benefit from the intervention and provided a majority of incorrect responses ('00' trials) in pre- and post-intervention blocks. Mirroring the overall accuracy effects, they represented 55.6% of reasoners in the control group but only 15.8% of reasoners in the

training group. Second, 'correct responders' provided a stable majority of correct answers ('01' or '11' trials) before and after the training intervention, and thus did not require any intervention to respond correctly. They represented 31.2% of reasoners in the training group and 22.7% in the control group. Third, 'improved responders' are those whose accuracy increased after the training intervention. They either gave a majority of biased responses ('00' trials) before the intervention and then switched to a majority of correct responses after the intervention ('01' or '11' trials), or already gave a majority of correct final responses ('01' trials) before the intervention but then switched to a majority of correct initial and final responses ('11' trials) after the intervention. They amounted to 50.5% of reasoners in the training group and 16.0% in the control group. Participants who gave inconsistent response patterns and could not be classified were put in the 'Other' category (2.5% in the training group, 5.8% in the control group; see Figure S4 in Supplementary Material Section J for full results).

No-conflict trial accuracy

For completeness, for each task and each participant, we also calculated the average proportion of initial 'intuitive' and final 'deliberate' responses for all the no-conflict items. Results showed that performance was consistently at ceiling in pre- and post-intervention blocks for initial responses (M = 90.5%, SD = 12.7 in the training group; M = 89.1%, SD = 11.6 in the control group), and final responses (M = 92.8%, SD = 11.2 in the training group; M = 90.5%, SD = 10.5 in the control group). The high initial and final performance on the no-conflict control problems provides evidence against a general systematic guessing confound (Bago & De Neys, 2017). In other words, if participants were not paying attention and were simply guessing throughout the study, they should have performed much worse on the no-conflict items. It also argues against a 'reversed heuristic' training account (Boissin et al., 2022) in which training would simply lead participants to distrust the intuitively cued response. If this were the case, we would expect a significant decline in post-intervention no-conflict trial performance (in which the intuitive, heuristic response was always correct). A detailed overview of the no-conflict problem accuracies by task can be found in Table S3 in Supplementary Material Section F.

General Discussion

In the present paper, we explored the debiasing effect of a short training that provides explanations for three different reasoning tasks (i.e., base-rate neglect, conjunction fallacy, and bat-and-ball tasks) in a French-speaking population. Especially, we explored whether we replicated previous debiasing findings observed in English-speaking populations (e.g., Boissin et al., 2021, 2022;

Claidière et al., 2017; Franiatte et al., 2024; Hoover & Healy, 2017; Morewedge et al., 2015; Purcell et al., 2020; Trouche et al., 2014). We used a two-response paradigm to track participants' initial 'intuitive' and final 'deliberate' responses.

Results indicated that the debiasing intervention led to a clear training effect. At the end of the session, a majority of trained reasoners were able to produce correct responses to structurally similar problems afterwards. Interestingly, the two-response findings indicated that this effect was observed as early as the initial 'intuitive' stage (overall 45% increase). That is, after training, reasoners no longer required correction of their erroneous intuitively generated heuristic response. Instead, they were able to produce intuitive responses consistent with logico-mathematical principles from the outset. In other words, the training intervention manages to get the majority of biased reasoners to intuit correctly. In this sense, the current study points to similar conclusions to previous debiasing studies conducted on English-speaking populations (e.g., Boissin et al., 2021, 2022; Franiatte et al., 2024; Hoover & Healy, 2017, Purcell et al., 2021).

Although we successfully replicated the training effect reported in the literature, there are also some notable differences between our study and previous debiasing findings that should be highlighted. A first thing to note is that participants in our study consistently outperformed Franiatte et al.'s (2024) reasoners in terms of accuracy. Specifically, in the French-speaking training group, before the intervention, we found a 10% higher initial accuracy and a 19% higher final accuracy compared to Franiatte et al.'s (2024) study. Similarly, after the intervention, we observed a 13% higher initial accuracy and a 7% higher final accuracy. These differences in performance could tentatively be attributed to the high levels of education reported by participants in the current study: 92% had a university degree, which is in contrast to previous studies where the proportion was around 50% (e.g., Boissin et al., 2021, 2022; Franiatte et al., 2024).

A second thing to note is that all participants in the current study were adult volunteers, whereas many previous psychological training studies have primarily involved college students or paid online workers (e.g., Prolific or MTurk; see Barrett, 2020). Consequently, the more highly educated volunteers in our study might have shown higher levels of motivation and engagement during the experiment. At the same time, the fact that despite these variations in sample composition, we still observed similar overall training effects underscores the robustness of the training. Nevertheless, it may be worthwhile in future work to examine how individual differences could potentially account for variations in training accuracy. Against this backdrop, numerous studies pointed to the fact that the accuracy of both initial and final answers can be affected by individual differences in, among others, cognitive abilities (e.g., intelligence), thinking styles (e.g., the propensity for reflection), cultural backgrounds, age or education (e.g., Boissin et al., 2023b; Srol & De Neys, 2021; Thompson &

Johnson, 2014; Raoelison et al., 2021). It may be worthwhile in future work to examine whether and how these individual differences factors impact training efficiency.

In this study, we found that the debiasing training effect can be generalized to a language and cultural context different from the English-speaking one. It is worth noting that comparing studies conducted in different languages presents challenges, as there is no way to ensure complete similarity in stimuli and instructions between languages (Boroditsky, 2001). Note, however, that it was not our primary objective to contrast the precise extent of the training effect per se in this study. Instead, we mainly wanted to test whether the training can successfully debias people's intuitive and deliberate reasoning in another language (i.e., whether there is a significant improvement to start with). Hence, empirically demonstrating that a debiasing training works, and impacts people's intuitive reasoning is far from trivial in this respect.

However, it is also clear that the approach we introduced here can be further developed. Hence there are a number of limitations that one needs to take in mind. First, we only focused on (native) French speakers, who were tested with problem contents designed for a French-speaking test environment. Although this was an initial step towards opening up the training to more diversity, it is important to note that these participants are still considered a WEIRD population (i.e., Western Educated Industrialized Rich and Democratic societies, see Henrich et al., 2010). Ideally, future studies should also investigate the debiasing effect on other languages, cultural groups, or populations who live in conditions vastly different than the current group of French-speaking Europeans (e.g., Boissin et al., 2024; Trémolière et al., 2022).

Second, a critic might argue that the observed improvement in trained reasoners could be solely attributed to the general feedback on correct and incorrect responses, rather than the training per se. However, as Janssen et al. (2020) showed with the bat-and-ball task, providing minimal feedback to reasoners, on average, does not significantly impact accuracy. One explanation for this lies in the nature of errors in these reasoning tasks. Notably, prior studies indicated that biased reasoners often show minimal error sensitivity or bias detection from the onset (see De Neys, 2023, for a review). That is, even without feedback, people seem to implicitly detect that their answer is not fully warranted. This tentatively indicates that people are not biased because they do not realize that their response is incorrect but rather because they do not explicitly know how to arrive at the correct solution strategy, thereby arguing against a general feedback confound. As Janssen et al. (2020) put it, a more informative retrieval cue would be needed to arrive at the correct solution. Hence, if we want people to reason more accurately, it appears necessary to provide additional information about the correct solution strategy beyond simple feedback.

Third, one may also wonder whether our training results simply stem from a mere repetition effect. As we presented 48 problems in a row, it cannot be excluded that some reasoners benefited from spontaneous learning simply by repeatedly solving structurally similar problems. For instance, some reasoners in the control group improved "naturally" (from pre- to post-intervention, initial responses rose by 6.4 points and final responses rose by 1.4 points). However, this improvement is marginal and our key interest is the effect of the debiasing intervention on reasoning performance - which is much more pronounced in the training than in the control group. Especially, the Block x Group interaction effect is significant, indicating that there is an effect of the training but not for the control group. If the repeated exposure had led to a strong spontaneous learning effect, we should have observed a non-significant Block x Group interaction. Additionally, note that previous studies such as Raoelison and De Neys (2019) investigated how repeated exposure affects initial and final response accuracy (on the bat-and-ball task), and showed that even extensive repeated exposure has a limited impact on reasoners' performance. Taken together, this seems to argue against a strong confounding spontaneous learning effect in our data.

Fourth, considering items of each task as random variables in our statistical model suggests that the training effect could readily generalize to other classic reasoning tasks. Although mastering these elementary logical principles is essential for sound reasoning, it's important to note that these lab-based tasks remain somewhat artificial (e.g., Janssen et al., 2021; Politzer et al., 2017; Prado et al., 2020). Arguably, as highlighted in the introduction, biased judgments can have detrimental impacts across various domains of everyday life. Therefore, it will remain important to test the transition from laboratory conditions to more ecological settings. One could think here, for example, of more applied context such as classroom settings (e.g., Brault Foisy et al., 2015), medical diagnosis (e.g., Topol, 2024) or gender discrimination in recruitment decisions (e.g., Isaac et al., 2009).

Fifth, note also that additional methodologies such as mouse tracking can be considered to better understand the dynamics of reasoning and decision-making (e.g., Freeman & Ambady, 2010; Spivey et al., 2005). In particular, hand movements provide a constant flow of data that can reveal ongoing dynamics of processing with fine-grained temporal sensitivity (Freeman et al., 2011). It could help to further pinpoint the time course of fast 'intuitive' responses (Travers et al., 2016), and provide deeper insights into the cognitive processes underlying training effects.

Finally, one may wonder whether the training effect is sustainable over time. Here, for mere practical reasons (we recruited volunteers and tested them without assigning an identifier), we did not investigate the robustness of the training effect. However, previous debiasing studies (e.g., Boissin et al., 2021, 2022) conducted a retest two months after the initial training session. Results indicated that the training effect persisted - at least after two months - with accuracies after two

months still being higher than before the initial training. Additionally, in a recent study (Franiatte et al., 2024), we found that the training effect could be boosted - and even made more robust over time - when participants take the training twice within a single week. Obviously, one could try to boost the training efficacy further with more immediate and/or frequent re-training. The optimal schedule remains to be explored here.

To conclude, in the present work, we replicated previous debiasing findings with a French-adapted training. This study suggests that simple interventions can be employed to boost sound reasoning – as early as the intuitive stage – in different parts of the world and confirms the suitability of the French versions for future research.

Data Accessibility Statement

Raw data, analysis scripts, and preregistration for this study can be downloaded from our OSF page (https://osf.io/hk8rv/).

Funding

This research was supported by a grant from the Agence Nationale de la Recherche (ANR-23-CE28-0004-01).

Acknowledgements

We would like to thank Erwan Le Bronec and Mégane Alekanian for their precious help in conducting this project. We would also like to thank Kenan Bouchentouf for his valuable work on the French conjunction fallacies items.

References

Andersson, L., Eriksson, J., Stillesjö, S., Juslin, P., Nyberg, L., & Wirebring, L. K. (2020). Neurocognitive processes underlying heuristic and normative probability judgments. *Cognition*, *196*, 104153. https://doi.org/10.1016/j.cognition.2019.104153

Arnett, J. J. (2008). The Neglected 95%: Why American Psychology Needs to Become Less American. *American Psychologist*, *63*(7), 602–614. https://doi.org/10.1037/0003-066X.63.7.602

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition, 158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The Smart System 1: evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*(3), 241–254. https://doi.org/10.1017/S0140525X07001653

Barrett, H. C. (2020). Towards a Cognitive Science of the Human: Cross-Cultural Approaches and Their Urgency. *Trends in Cognitive Science, 2(8),* 620–638. https://doi.org/10.1016/j.tics.2020.05.007

Bender, A., & Beller, S. (2019). The Cultural Fabric of Human Causal Cognition. *Perspectives on Psychological Science*, *14*(6), 922–940. https://doi.org/10.1177/1745691619863055

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Science, 26(12)*, 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Boissin, E., Caparos, S., Raoelison, M., & De Neys, W. (2021). From bias to sound intuiting: Boosting correct intuitive reasoning. *Cognition*, *211,* 104645. https://doi.org/10.1016/j.cognition.2021.104645

Boissin, E., Caparos, S., Voudouri, A., & De Neys, W. (2022). Debiasing System 1: Training favours logical over stereotypical intuiting. *Judgment and Decision Making*, *17*(4), 646–690. https://doi.org/10.1017/S1930297500008895

Boissin, E., Caparos, S., De Neys W. (2023a). No easy fix for belief bias during syllogistic reasoning?. *Journal of Cognitive Psychology, 35(4)*. https://doi.org/10.1080/20445911.2023.2181734

Boissin, E., Caparos, S., Borst, G., & De Neys, W. (2023b). Debiasing intuitive thinking is less effective in younger than older adolescents: The role of knowledge instantiation. [Manuscript in preparation].

Boissin, E., Josserand, M., De Neys, W., & Caparos, S. (2024). Debiasing thinking among non-WEIRD reasoners. *Cognition, 243*, 105681. https://doi.org/10.1016/j.cognition.2023.105681

Boroditsky, L. (2001). Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive psychology, 43(1)*, 1-22. https://doi.org/10.1006/cogp.2001.0748

Brault Foisy, L.-M., Ahr, E., Masson, S., Borst, G., & Houdé, O. (2015). Blocking our brain: When we need to inhibit repetitive mistakes! *Frontiers for Young Minds, 5*. https://doi.org/10.3389/frym.2015.00017

Brysbaert, M., & Debeer, D. (2023). How to run linear mixed effects analysis for pairwise comparisons? A tutorial and a proposal for the calculation of standardized effect sizes. https://doi.org/10.31234/osf.io/esnku

Claidière, N., Trouche, E., & Mercier, H. (2017). Argumentation and the diffusion of counter-intuitive beliefs. *Journal of Experimental Psychology: General*, *146*(7), 1052–1066. https://doi.org/10.1037/xge0000323

De Neys, W. (2006). Automatic-heuristic and executive-analytic processing during reasoning: Chronometric and dual-task considerations. *Quarterly Journal of Experimental Psychology*, *59*(6), 1070–1100. https://doi.org/10.1080/02724980543000123

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking and Reasoning*, *20*(2), 169–187. https://doi.org/10.1080/13546783.2013.854725

De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 1–68. https://doi.org/10.1017/S0140525X2200142X

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, *6*(1), e15954. https://doi.org/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 and Review*, *20*(2), 269–273. https://doi.org/10.3758/s13423-013-0384-5

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7(10)*, 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59(1)*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. S. B. T. (2010). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry*, *21*(4), 313–326. https://doi.org/10.1080/1047840X.2010.521057

Evans, J. S. B. T. (2019). Reflections on reflection: the nature and function of type 2 processes in dual-process theories of reasoning. *Thinking and Reasoning*, *25(4)*, 383–415. https://doi.org/10.1080/13546783.2019.1623071

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448. https://doi.org/10.1017/S0140525X0999094X

Franiatte, N., Boissin, E., Delmas, A., & De Neys, W. (2024). Boosting debiasing: Impact of repeated training on reasoning. *Learning and Instruction, 89*, 101845. https://doi.org/10.1016/j.learninstruc.2023.101845

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior research methods, 42(1)*, 226-241. https://doi.org/10.3758/BRM.42.1.226

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in psychology, 2*, 59. https://doi.org/10.3389/fpsyg.2011.00059

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking and Reasoning*, *15*(2), 105–128. https://doi.org/10.1080/13546780802711185

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and Content: The Use of Base Rates as a Continuous Variable. *Journal of Experimental Psychology: Human Perception and Performance*, *14(3),* 513-525. https://doi.org/10.1037/0096-1523.14.3.513

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/https://doi.org/10.1017/S0140525X0999152X

Hoover, J. D., & Healy, A. F. (2017). Algebraic reasoning and bat-and-ball problem variants: Solving isomorphic algebra first facilitates problem solving later. *Psychonomic Bulletin and Review*, *24*(6), 1922–1928. https://doi.org/10.3758/s13423-017-1241-8

Huettig, F., & Ferreira, F. (2023). The Myth of Normal Reading. *Perspectives on Psychological Science*, *18(4),* 863–870. https://doi.org/10.1177/17456916221127226

Isaac, C., Lee, B., & Carnes, M. (2009). Interventions that affect gender bias in hiring: A systematic review. *Academic Medicine: Journal of the Association of American Medical Colleges, 84(10)*, 1440. https://doi.org/10.1097/ACM.0b013e3181b6ba00

Janssen, E. M., Raoelison, M., & De Neys, W. (2020). "You're wrong!": The impact of accuracy feedback on the bat-and-ball problem. *Acta psychologica, 206*, 103042. https://doi.org/10.1016/j.actpsy.2020.103042

Janssen, E. M., Velinga, S. B., De Neys, W., & Van Gog, T. (2021). Recognizing biased reasoning: Conflict detection during decision-making and decision-evaluation. *Acta Psychologica, 217*, Article 103322. https://doi.org/10.1016/j.actpsy.2021.103322

Kahneman, D. (2011). *Thinking, fast and slow.* New York, NY: Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In & R. G. M. (Eds.), in K. J. Holyoak (Ed.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review, 80(4)*, 237. https://doi.org/10.1037/h0034747

Kuznetsova, A., Brockhoff, PB., Christensen, RHB. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82(13)*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare?. *Perspectives on psychological science, 4(4)*, 390-398. https://doi.org/10.1111/j.1745-6924.2009.01144.x

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved?. *Perspectives on psychological science, 4(4)*, 379-383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621–640. https://doi.org/10.1037/0096-3445.130.4.621

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in cognitive sciences, 14(10)*, 435-440. https://doi.org/10.1016/j.tics.2010.07.004

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing Decisions: Improved Decision Making With a Single Training Intervention. *Policy*

Insights from the Behavioral and Brain Sciences, *2*(1), 129–140. https://doi.org/10.1177/2372732215600886

Nisbett, R. E. (1993). *Rules for reasoning.* Psychology Press. https://doi.org/10.4324/9780203763230

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(2), 544–554. https://doi.org/10.1037/a0034887

Politzer, G., Bosc-Miné, C., & Sander, E. (2017). Preadolescents solve natural syllogisms proficiently. *Cognitive Science, 41(S5)*, 1031–1061. https://doi.org/10.1111/cogs.12365

Prado, J., Léone, J., Epinat-Duclos, J., Trouche, E., & Mercier, H. (2020). The neural bases of argumentative reasoning. *Brain and Language, 208*, Article 104827. https://doi.org/10.1016/j.bandl.2020.104827

Purcell, Z. A., Wastell, C. A., & Sweller, N. (2020). Domain-specific experience and dual-process thinking. *Thinking and Reasoning*, *27(2)*, 1–29. https://doi.org/10.1080/13546783.2020.1793813

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision Making, 14(2)*, 170-178. https://doi.org/10.1017/S1930297500003405

Raoelison, M., Boissin, E., Borst, G., & De Neys, W. (2021). From slow to fast logic: the development of logical intuitions. *Thinking & Reasoning*, *27*(4), 599–622. https://doi.org/10.1080/13546783

Smith, L., Leung, W. G., Crane, B., Parkinson, B., Toulopoulou, T., & Yiend, J. (2018). Bilingual comparison of Mandarin and English cognitive bias tasks. *Behavior Research Methods*, *50*(1), 302–312. https://doi.org/10.3758/s13428-017-0871-0

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences, 102*, 10393-10398. https://doi.org/10.1073/pnas.0503903102

Šrol, J., & De Neys, W. (2021). Predicting individual differences in conflict detection and bias susceptibility during reasoning. *Thinking & Reasoning, 27(1)*, 38-68. https://doi.org/10.1080/13546783.2019.1708793

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behavioral and Brain Sciences, 23(5)*, 645–665. https://doi.org/10.1017/S0140525X00003435

Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2021). The neglected 95% revisited: Is American psychology becoming less American? American Psychologist, 76(1), 116–129. https://doi.org/10.1037/amp0000622

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning, 20(2)*, 215-244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Topol, E. J. (2024). Toward the eradication of medical diagnostic errors. *Science, 383(6681)*, eadn9602. https://doi.org/10.1126/science.adn9602

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. *Cognition, 150*, 109-118. https://doi.org/10.1016/j.cognition.2016.01.015

Trémolière, B., Davidoff, J. B., & Caparos, S. (2022). A 21st century cognitive portrait of the Himba, a remote people of Namibia. *British Journal of Psychology*, *113*(2), 508–530. https://doi.org/10.1111/bjop.12539

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971. https://doi.org/10.1037/a0037099

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review, 90(4)*, 293-315. https://doi.org/10.1037/0033-295X.90.4.293

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., Müller, K. (2020). dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr

Wickham, H., Vaughan, D., Girlich, M. (2024). tidyr: Tidy Messy Data. R package version 1.3.1, https://github.com/tidyverse/tidyr, https://tidyr.tidyverse.org

Supplementary Material

A. Material: French items used in the study

BR = Base-Rate neglect, CF = Conjunction Fallacy, and BB = Bat-and-Ball tasks.

	Task	Conflict version	No-conflict version
1	BR	Cette étude concerne des boxeuses et des caissières de supermarché. La personne 'W' est musclée. Il y a 5 boxeuses et 995 caissières de supermarché. Est-ce que la personne 'W' a plus de chance d'être : • Une boxeuse • Une caissière de supermarché	Cette étude concerne des boxeuses et des caissières de supermarché. La personne 'W' est musclée. Il y a 995 boxeuses et 5 caissières de supermarché. Est-ce que la personne 'W' a plus de chance d'être : • Une boxeuse • Une caissière de supermarché
2	BR	Cette étude concerne des architectes et des chauffeurs de bus. La personne 'C' est créative. Il y a 6 architectes et 994 chauffeurs de bus. Est-ce que la personne 'C' a plus de chance d'être : • Un architecte • Un chauffeur de bus	Cette étude concerne des architectes et des chauffeurs de bus. La personne 'C' est créative. Il y a 994 architectes et 6 chauffeurs de bus. Est-ce que la personne 'C' a plus de chance d'être : • Un architecte • Un chauffeur de bus
3	BR	Cette étude concerne des écrivains et des ouvriers de chantier. La personne 'F' est robuste. Il y a 996 écrivains et 4 ouvriers de chantier. Est-ce que la personne 'F' a plus de chance d'être : • Un écrivain • Un ouvrier de chantier	Cette étude concerne des écrivains et des ouvriers de chantier. La personne 'F' est robuste. Il y a 4 écrivains et 996 ouvriers de chantier. Est-ce que la personne 'F' a plus de chance d'être : • Un écrivain • Un ouvrier de chantier
4	BR	Cette étude concerne des directeurs administratifs et des humoristes. La personne 'K' est drôle. Il y a 997 directeurs administratifs et 3 humoristes. Est-ce que la personne 'K' a plus de chance d'être : • Un directeur administratif • Un humoriste	Cette étude concerne des directeurs administratifs et des humoristes. La personne 'K' est drôle. Il y a 3 directeurs administratifs et 997 humoristes. Est-ce que la personne 'K' a plus de chance d'être : • Un directeur administratif • Un humoriste
5	BR	Cette étude concerne des hôtesses de l'air et des gardiens de prison. La personne 'M' est charmante. Il y a 3 hôtesses de l'air et 997 gardiens de prison. Est-ce que la personne 'M' a plus de chance d'être : • Une hôtesse de l'air • Un gardien de prison	Cette étude concerne des hôtesses de l'air et des gardiens de prison. La personne 'M' est charmante. Il y a 997 hôtesses de l'air et 3 gardiens de prison. Est-ce que la personne 'M' a plus de chance d'être : • Une hôtesse de l'air • Un gardien de prison
6	BR	Cette étude concerne des pompiers et des riches héritiers. La personne 'L' est courageuse. Il y a 4 pompiers et 996 riches héritiers. Est-ce que la personne 'L' a plus de chance d'être : • Un pompier • Un riche héritier	Cette étude concerne des pompiers et des riches héritiers. La personne 'L' est courageuse. Il y a 996 pompiers et 4 riches héritiers. Est-ce que la personne 'L' a plus de chance d'être : • Un pompier • Un riche héritier
7	BR	Cette étude concerne des éboueurs et des hommes d'affaires. La personne 'D' est ambitieuse. Il y a 994 éboueurs et 6 hommes d'affaires. Est-ce que la personne 'D' a plus de chance d'être :	Cette étude concerne des éboueurs et des hommes d'affaires. La personne 'D' est ambitieuse. Il y a 6 éboueurs et 994 hommes d'affaires. Est-ce que la personne 'D' a plus de chance d'être :

		• Un éboueur • Un homme d'affaire	• Un éboueur • Un homme d'affaire
8	BR	Cette étude concerne des jardiniers et des PDG. La personne 'S' est autoritaire. Il y a 995 jardiniers et 5 PDG. Est-ce que la personne 'S' a plus de chance d'être : • Un jardinier • Un PDG	Cette étude concerne des jardiniers et des PDG. La personne 'S' est autoritaire. Il y a 5 jardiniers et 995 PDG. Est-ce que la personne 'S' a plus de chance d'être : • Un jardinier • Un PDG
9	BR	Cette étude concerne des agents immobiliers et des chômeurs. La personne 'X' est illettrée. Il y a 994 agents immobiliers et 6 chômeurs. Est-ce que la personne 'X' a plus de chance d'être : • Un agent immobilier • Un chômeur	Cette étude concerne des agents immobiliers et des chômeurs. La personne 'X' est illettrée. Il y a 6 agents immobiliers et 994 chômeurs. Est-ce que la personne 'X' a plus de chance d'être : • Un agent immobilier • Un chômeur
10	BR	Cette étude concerne des chirurgiennes et des adolescentes. La personne 'V' est immature. Il y a 995 chirurgiennes et 5 adolescentes. Est-ce que la personne 'V' a plus de chance d'être : • Une chirurgienne • Une adolescente	Cette étude concerne des chirurgiennes et des adolescentes. La personne 'V' est immature. Il y a 5 chirurgiennes et 995 adolescentes. Est-ce que la personne 'V' a plus de chance d'être : • Une chirurgienne • Une adolescente
11	BR	Cette étude concerne des dentistes et des profs de sport. La personne 'J' est méticuleuse. Il y a 3 dentistes et 997 profs de sport. Est-ce que la personne 'J' a plus de chance d'être : • Un dentiste • Un prof de sport	Cette étude concerne des dentistes et des profs de sport. La personne 'J' est méticuleuse. Il y a 997 dentistes et 3 profs de sport. Est-ce que la personne 'J' a plus de chance d'être : • Un dentiste • Un prof de sport
12	BR	Cette étude concerne des bibliothécaires et des DJ. La personne 'R' est calme. Il y a 5 bibliothécaires et 995 DJ. Est-ce que la personne 'R' a plus de chance d'être : • Un bibliothécaire • Un DJ	Cette étude concerne des bibliothécaires et des DJ. La personne 'R' est calme. Il y a 995 bibliothécaires et 5 DJ. Est-ce que la personne 'R' a plus de chance d'être : • Un bibliothécaire Un DJ
13	BR	Cette étude concerne des nourrices et des femmes d'affaires. La personne 'H' est attentionnée. Il y a 4 nourrices et 996 femmes d'affaires. Est-ce que la personne 'H' a plus de chance d'être : • Une nourrice Une femme d'affaire	Cette étude concerne des nourrices et des femmes d'affaires. La personne 'H' est attentionnée. Il y a 996 nourrices et 4 femmes d'affaires. Est-ce que la personne 'H' a plus de chance d'être : • Une nourrice • Une femme d'affaire
14	BR	Cette étude concerne des scientifiques et des vendeuses par téléphone. La personne 'E' est rigoureuse. Il y a 6 scientifiques et 994 vendeuses par téléphone. Est-ce que la personne 'E' a plus de chance d'être : • Un scientifique • Une vendeuse de téléphone	Cette étude concerne des scientifiques et des vendeuses par téléphone. La personne 'E' est rigoureuse. Il y a 994 scientifiques et 6 vendeuses par téléphone. Est-ce que la personne 'E' a plus de chance d'être : • Un scientifique • Une vendeuse de téléphone
15	BR	Cette étude concerne des juges et des secrétaires. La personne 'T' est à l'écoute. Il y a 996 juges et 4 secrétaires. Est-ce que la personne 'T' a plus de chance d'être : • Un juge • Une secrétaire	Cette étude concerne des juges et des secrétaires. La personne 'T' est à l'écoute. Il y a 4 juges et 996 secrétaires. Est-ce que la personne 'T' a plus de chance d'être : • Un juge • Une secrétaire

16	BR	Cette étude concerne des procureures et des infirmières. La personne 'U' est rassurante. Il y a 997 procureures et 3 infirmières. Est-ce que la personne 'U' a plus de chance d'être : • Une procureure • Une infirmière	Cette étude concerne des procureures et des infirmières. La personne 'U' est rassurante. Il y a 3 procureures et 997 infirmières. Est-ce que la personne 'U' a plus de chance d'être : • Une procureure • Une infirmière
17	CF	Serge, 25 ans, a étudié l'aérodynamique et aime les sports extrêmes. Est-il plus probable que la personne décrite soit : • Professeur d'histoire • Croque-mort • Professeur d'histoire et joueur de scrabble • Professeur d'histoire et pilote de moto	Clara, 45 ans, a étudié l'aérodynamique et aime les sports extrêmes. Est-il plus probable que la personne décrite soit : • Pilote de moto • Croque-mort • Professeur d'histoire et joueur de scrabble • Professeur d'histoire et pilote de moto
18	CF	Clara, 26 ans, a étudié le marketing web et aime les réseaux sociaux. Est-il plus probable que la personne décrite soit : • Gendarme • Avaleur d'épée • Gendarme et fan de puzzles • Gendarme et youtubeur	Serge, 44 ans, a étudié le marketing web et aime les réseaux sociaux. Est-il plus probable que la personne décrite soit : • Youtubeur • Avaleur d'épée • Gendarme et fan de puzzles • Gendarme et youtubeur
19	CF	Camille, 27 ans, a étudié la robotique et aime les Intelligences Artificielles. Est-il plus probable que la personne décrite soit : • Caissier • Chanteur de pop international • Caissier et cheerleader • Caissier et hackeur	Lucas, 43 ans, a étudié la robotique et aime les Intelligences Artificielles. Est-il plus probable que la personne décrite soit : • Hackeur • Chanteur de pop international • Caissier et cheerleader • Caissier et hackeur
20	CF	Lucas, 29 ans, a étudié la psychologie et aime les œuvres caritatives. Est-il plus probable que la personne décrite soit : • Huissier • Charmeur de serpent • Huissier et parieur sportif • Huissier et bénévole	Camille, 41 ans, a étudié la psychologie et aime les œuvres caritatives. Est-il plus probable que la personne décrite soit : • Bénévole • Charmeur de serpent • Huissier et parieur sportif • Huissier et bénévole
21	CF	Charles, 35 ans, a étudié la philosophie et aime la Grèce antique. Est-il plus probable que la personne décrite soit : • Coach sportif • Eleveur d'otarie • Coach sportif et fan de télé-réalité • Coach sportif et collectionneur d'art	Chloé, 36 ans, a étudié la philosophie et aime la Grèce antique. Est-il plus probable que la personne décrite soit : • Collectionneur d'art • Eleveur d'otarie • Coach sportif et fan de télé-réalité • Coach sportif et collectionneur d'art
22	CF	Mathieu, 39 ans, a étudié la comédie et aime rire. Est-il plus probable que la personne décrite soit : • Archiviste • Directeur de banque • Archiviste et karatéka • Archiviste et clown	Manon, 33 ans, a étudié la comédie et aime rire. Est-il plus probable que la personne décrite soit : • Clown • Directeur de banque • Archiviste et karatéka • Archiviste et clown
23	CF	Manon, 40 ans, a étudié l'immobilier et aime les objets de luxe. Est-il plus probable que la personne décrite soit : • Livreur de pizza • Capitaine de sous-marin • Livreur de pizza et maquilleur	Mathieu, 32ans, a étudié l'immobilier et aime les objets de luxe. Est-il plus probable que la personne décrite soit : • Un collectionneur de montres • Capitaine de sous-marin • Livreur de pizza et maquilleur

		• Livreur de pizza et collectionneur de montres	• Livreur de pizza et collectionneur de montres
24	CF	David, 41 ans, a étudié l'art du cirque et aime la gymnastique. Est-il plus probable que la personne décrite soit : • Ramasseur de fruits • Chef d'état • Ramasseur de fruits et joueur de jeux vidéo • Ramasseur de fruits et acrobate	Lucie, 31 ans, a étudié l'art du cirque et aime la gymnastique. Est-il plus probable que la personne décrite soit : • Acrobate • Chef d'état • Ramasseur de fruits et joueur de jeux vidéo • Ramasseur de fruits et acrobate
25	CF	Steeve, 33 ans, a étudié la biologie et aime les balades en forêt. Est-il plus probable que la personne décrite soit : • Masseur • Pilote de chasse • Masseur et lutteur • Masseur et ramasseur de champignon	Amélie, 37 ans, a étudié la biologie et aime les balades en forêt. Est-il plus probable que la personne décrite soit : • Ramasseur de champignon • Pilote de chasse • Masseur et lutteur • Masseur et ramasseur de champignon
26	CF	Amélie, 32 ans, a étudié la théologie et aime les chants de chorale. Est-il plus probable que la personne décrite soit : • Travailleur dans un entrepôt • Conducteur de Formule 1 • Travailleur dans un entrepôt et joueur de paintball • Travailleur dans un entrepôt et chrétien	Steeve, 38 ans, a étudié la théologie et aime les chants de chorale. Est-il plus probable que la personne décrite soit : • Chrétien • Conducteur de Formule 1 • Travailleur dans un entrepôt et joueur de paintball • Travailleur dans un entrepôt et chrétien
27	CF	Antoine, 31 ans, a étudié l'informatique et aime les mangas. Est-il plus probable que la personne décrite soit : • Barman • Diplomate • Barman et fumeur de pipe • Barman et joueur en ligne	Sophia, 39 ans, a étudié l'informatique et aime les mangas. Est-il plus probable que la personne décrite soit : • Joueur en ligne • Diplomate • Barman et fumeur de pipe • Barman et joueur en ligne
28	CF	Sophia, 30 ans, a étudié l'économie et aime le bon tabac. Est-il plus probable que la personne décrite soit : • Assistant de vente • Snowboardeur professionnel • Assistant de vente et danseur de ballet • Assistant de vente et fumeur de cigare	Antoine, 40 ans, a étudié l'économie et aime le bon tabac. Est-il plus probable que la personne décrite soit : • Fumeur de cigare • Snowboardeur professionnel • Assistant de vente et danseur de ballet • Assistant de vente et fumeur de cigare
29	CF	Didier, 29 ans, a étudié l'ingénierie du son et aime les chaines hifis. Est-il plus probable que la personne décrite soit : • Boulanger • Comte • Boulanger et pratiquant de sports extrêmes • Boulanger et fan de musique	Adèle, 29 ans, a étudié l'ingénierie du son et aime les chaines hifis. Est-il plus probable que la personne décrite soit : • Fan de musique • Comte • Boulanger et pratiquant de sports extrêmes • Boulanger et fan de musique
30	CF	Adèle, 27 ans, a étudié le stylisme et aime la couture. Est-il plus probable que la personne décrite soit : • Aide-soignant • Astronaute • Aide-soignant et généalogiste • Aide-soignant et passionné de mode	Didier, 43 ans, a étudié le stylisme et aime la couture. Est-il plus probable que la personne décrite soit : • Passionné de mode • Astronaute • Aide-soignant et généalogiste • Aide-soignant et passionné de mode
31	CF	Samuel, 26 ans, a étudié les sciences de l'éducation et aime les enfants. Est-il plus probable que la personne décrite soit :	Nelson, 44 ans, a étudié les sciences de l'éducation et aime les enfants. Est-il plus probable que la personne décrite soit :

		• Agent de bord • Duc • Agent de bord et fan de courses de rallye • Agent de bord et père au foyer	• Père au foyer • Duc • Agent de bord et fan de courses de rallye • Agent de bord et père au foyer
32	CF	Julie, 34 ans, a étudié le féminisme et aime la musique hardcore. Est-il plus probable que la personne décrite soit : • Cordonnier • Archevêque • Cordonnier et témoin de Jéhovah • Cordonnier et féministe	Nicolas, 35 ans, a étudié le féminisme et aime la musique hardcore. Est-il plus probable que la personne décrite soit : • Féministe • Archevêque • Cordonnier et témoin de Jéhovah • Cordonnier et féministe
33	BB	Dans un supermarché, on peut acheter 320 tomates et avocats. Il y a 300 tomates de plus que d'avocats. Combien y a-t-il d'avocats ?	Dans un supermarché, on peut acheter 160 tomates et avocats. Il y a 100 tomates. Combien y a-t-il d'avocats dans ce supermarché ?
34	BB	Dans une cuisine, il y a 260 couteaux et cuillères au total. Il y a 200 couteaux de plus que de cuillères. Combien y a-t-il de cuillères ?	Dans une cuisine, il y a 220 couteaux et cuillères au total. Il y a 200 couteaux. Combien y a-t-il de cuillères dans la cuisine ?
35	BB	Un magasin de musique a 210 saxophones et flûtes au total. Il y a 200 saxophones de plus que de flûtes. Combien y a-t-il de flûtes ?	Un magasin de musique a 270 saxophones et flûtes au total. Il y a 200 saxophones. Combien y a-t-il de flûtes dans le magasin ?
36	BB	Dans une entreprise, il y a 150 hommes et femmes au total. Il y a 100 hommes de plus que de femmes. Combien y a-t-il de femmes ?	Dans une entreprise, il y a 330 hommes et femmes au total Il y a 300 hommes. Combien y a-t-il de femmes dans l'entreprise ?
37	BB	Un parc national a 650 roses et orchidées au total. Il y a 600 roses de plus que d'orchidées. Combien y a-t-il d'orchidées ?	Un parc national a 380 roses et orchidées au total. Il y a 300 roses. Combien y a-t-il d'orchidées dans le parc ?
38	BB	Dans une piscine, il y a 540 nageurs et plongeurs au total. Il y a 500 nageurs de plus que de plongeurs. Combien y a-t-il de plongeurs ?	Dans une piscine, il y a 490 nageurs et plongeurs au total. Il y a 400 nageurs. Combien y a-t-il de plongeurs dans cette piscine ?
39	BB	Dans un magasin, il y a 480 clous et marteaux au total. Il y a 400 clous de plus que de marteaux. Combien y a-t-il de marteaux dans le magasin ?	Dans un magasin, il y a 550 clous et marteaux au total. Il y a 500 clous. Combien y a-t-il de marteaux dans le magasin ?
40	BB	Une ville possède 430 bus et trains au total. Il y a 400 bus de plus que de trains. Combien y a-t-il de trains dans la ville ?	Une ville possède 610 bus et trains au total. Il y a 600 bus. Combien y a-t-il de trains dans la ville ?
41	BB	Dans une forêt, il y a 640 chênes et érables au total. Il y a 600 chênes de plus que d'érables. Combien y a-t-il d'érables ?	Dans une forêt, il y a 390 chênes et érables au total. Il y a 300 chênes. Combien y a-t-il d'érables dans cette forêt ?
42	BB	Une entreprise emploie 580 techniciens et ingénieurs au total. Il y a 500 techniciens de plus que d'ingénieurs. Combien y a-t-il d'ingénieurs ?	Une entreprise emploie 450 techniciens et ingénieurs au total. Il y a 400 techniciens. Combien y a-t-il d'ingénieurs dans cette entreprise ?
43	BB	Pour un tournoi sportif, on a invité 530 joueurs et entraîneurs. Il y a 500 joueurs de plus que d'entraîneurs. Combien y a-t-il d'entraîneurs ?	Pour un tournoi sportif, on a invité 510 joueurs et entraîneurs. Il y a 500 joueurs. Combien y a-t-il d'entraîneurs invités à ce tournoi ?
44	BB	Sur une étagère, il y a 560 vis et tournevis au total. Il y a 500 vis de plus que de tournevis. Combien y a-t-il de tournevis sur l'étagère ?	Sur une étagère, il y a 560 vis et tournevis au total. Il y a 500 vis. Combien y a-t-il de tournevis sur l'étagère ?
45	BB	Un directeur de magasin a acheté 310 bananes et kiwis.	Un directeur de magasin a acheté 170 bananes et kiwis.

		Il y a 300 bananes de plus que de kiwis. Combien y a-t-il de kiwis ?	Il y a 100 bananes. Combien y a-t-il de kiwis dans ce magasin ?
46	BB	Dans un restaurant, il y a 250 verres et tasses au total. Il y a 200 verres de plus que de tasses. Combien y a-t-il de tasses ?	Dans un restaurant, il y a 230 verres et tasses au total. Il y a 200 verres. Combien y a-t-il de tasses dans ce restaurant ?
47	BB	Un magasin met en exposition 190 pianos et harpes. Il y a 100 pianos de plus que de harpes. Combien y a-t-il de harpes ?	Un magasin met en exposition 280 pianos et harpes. Il y a 200 pianos. Combien y a-t-il de harpes dans ce magasin ?
48	BB	Dans un parc, il y a 140 adultes et enfants au total. Il y a 100 adultes de plus que d'enfants. Combien y a-t-il d'enfants ?	Dans un parc, il y a 340 adultes et enfants au total. Il y a 300 adultes. Combien y a-t-il d'enfants ?

Note: For the base-rate task, we labelled the response that is in line with the base rates as the correct response. Critics of the base rate task (e.g., Barbey & Sloman, 2007; Gigerenzer et al., 1988) have long pointed out that if reasoners adopt a Bayesian approach and combine the base rate probabilities with the stereotypical description, this can lead to interpretative complications when the description is extremely diagnostic. For example, imagine that we have an item with males and females as the two groups and give the description that Person 'A' 'is 'pregnant'. Now, in this case, one would always need to conclude that Person 'A' is a woman, regardless of the base rates. The more moderate descriptions (such as 'kind' or 'creative') help to avoid this potential problem. In addition, the extreme base rates (i.e., 997/3, 996/4, 995/5) that were used in the current study further help to guarantee that even a very approximate Bayesian reasoner would need to pick the response cued by the base rates (see De Neys, 2014).

B. Conjunction fallacy problems: Frequency of each individual response option on conflict items

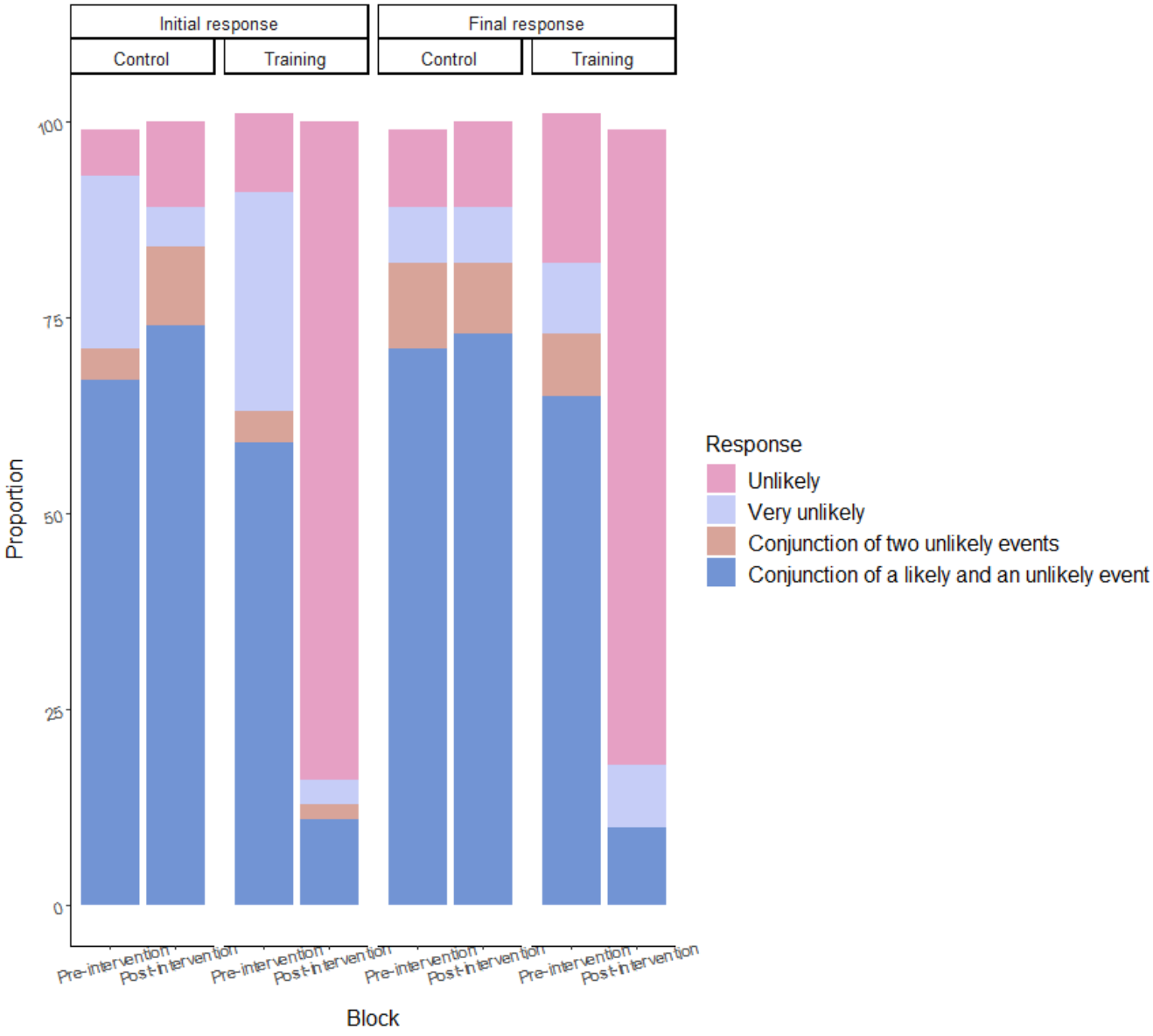


Figure S1. Frequency of each individual response option (conjunction fallacy conflict items) for the initial and the final responses, before and after the intervention in the control and training group.

C. Justification data

Table S1.

Frequency of different types of justifications for the final bat-and-ball (BB), base-rate (BR), conjunction fallacy (CF) conflict problems and all tasks combined (All) during the post-intervention of the study.

Task	Justification types	Control group		Training group	
		Correct response (n = 96)	*Incorrect response (n = 129)*	*Correct response (n = 170)*	*Incorrect response (n = 31)*
All	Math - Correct	64	3	105	2
	Math – Incorrect/Unspecified	7	45	4	9
	Guess	1	6	1	2
	Intuitions	13	53	30	13
	Other	11	22	30	5
BB	Math - Correct	35	3	43	-
	Math – Incorrect/Unspecified	-	18	-	6
	Guess	1	1	-	-
	Intuitions	3	8	7	5
	Other	-	3	4	2
BR	Math - Correct	25	-	37	1
	Math – Incorrect/Unspecified	7	4	1	1
	Guess	-	3	-	1
	Intuitions	9	13	8	3
	Other	10	6	14	-
CF	Math - Correct	4	-	25	1
	Math – Incorrect/Unspecified	-	23	3	2
	Guess	-	2	1	1
	Intuitions	1	32	15	5
	Other	1	13	12	3

Note: The coding format and procedure were based on Bago and De Neys (2019) for bat-and-ball, Boissin et al. (2022) for base-rate, and Franiatte et al. (2024) for conjunction fallacy tasks. A justification was considered correct when it explicitly mentioned the correct calculation for the bat-and-ball (e.g., '140 in total - 100 adults = 40 children / 2, the response is 20') or the use of the base-rate (e.g., 'Greater number of writers to constructions workers. For every 1 construction worker there are 249 writers, so the odds are stacked against it being a writer') or when it explicitly referred to the conjunction principle (e.g., 'There are always more people who are simply longshoreman than longshoreman and stargazer'). Other justifications, whether they mentioned an incorrect calculation or unspecified statement (e.g., 'I did it in my head') were coded as incorrect.

D. Confidence data

Note that due to a coding error the confidence data was not systematically recorded. The non-missing data is included in our data file on OSF, and an exploratory analysis of the partial data can be found below. We explored whether the training intervention affected biased reasoners' ability to detect conflict (i.e., conflict detection).

Table S2.

Conflict detection results on non-missing data. Percentage of mean difference in confidence ratings (SD) between initial correct no-conflict and initial incorrect conflict problems on each reasoning task: Bat-and-ball (BB), base-rate neglect (BR) and conjunction fallacy (CF).

Task	Group	Initial response - Session 1	
		Pre-intervention	Post-intervention
BB	Control	10.4 (24.8)	4.2 (19.9)
	Training	5.0 (23.3)	17.4 (33.3)
BR	Control	4.6 (21.4)	7.4 (17.4)
	Training	15.6 (36.8)	22.8 (32.7)
CF	Control	8.8 (16.2)	1.7 (20.8)
	Training	3.8 (13.8)	20.0 (27.4)

E. Bat-and-ball problems: Accuracy with and without reasoners who already knew the original bat-and-ball problem

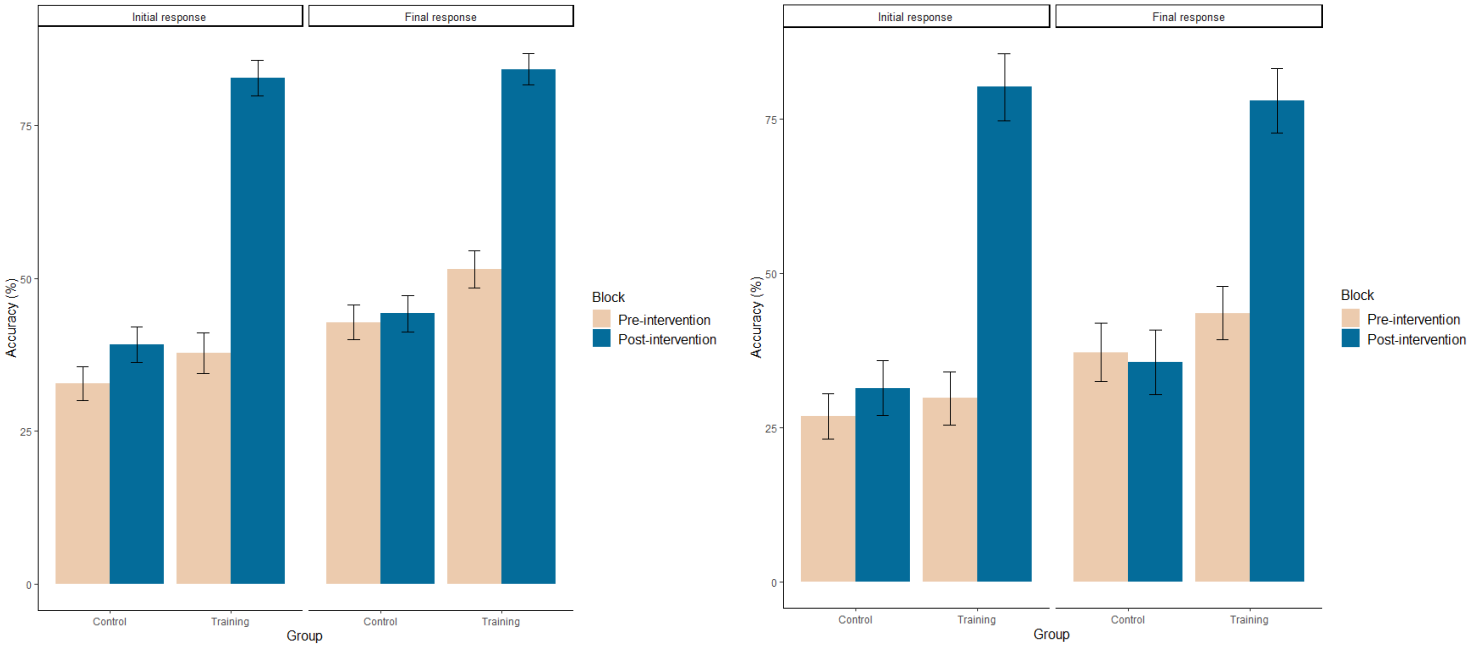


***Figure S2*.** Mean accuracy (%) of correct initial and final responses on conflict problems before and after the intervention, with (left panel) and without (right panel) reasoners who already knew the original bat-and-ball problem (Frederick, 2005). Error bars are standard errors. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks.

F. Accuracy for no-conflict problems

Table S3.

Average accuracy (%) for the no-conflict problems (SD) of bat-and-ball (BB), base-rate (BR) and conjunction fallacy (CF) tasks and combined (All task).

Task	Group	Initial response		Final response	
		Pre-intervention	Post-intervention	Pre-intervention	Post-intervention
BB	Control	78.8 (22.4)	96.1 (14.9)	82.6 (20.7)	97.4 (13.2)
	Training	79.5 (24.8)	91.0 (25.1)	84.6 (20.2)	93.5 (23.7)
BR	Control	95.7 (14.8)	94.0 (16.7)	98.3 (6.7)	95.4 (12.3)
	Training	87.9 (27.4)	95.3 (15.4)	93.6 (21.3)	96.6 (12.5)
CF	Control	83.8 (22.8)	73.4 (25.0)	85.6 (18.9)	73.1 (27.2)
	Training	86.7 (18.6)	94.0 (14.7)	86.0 (21.6)	93.2 (16.1)
All task	Control	85.9 (11.7)	87.7 (10.8)	88.3 (9.8)	88.5 (11.1)
	Training	84.3 (14.2)	93.5 (11.3)	87.7 (12.6)	94.4 (10.8)

G. Inferential statistics

Table S4.

Wald test results for the initial conflict responses, assessing the statistical significance of the fixed effect of the model with a 95% confidence interval. The significant p-values ($p < .05$) are marked with a *.

	Chi Square	Df	*p*
Group	1.43	1	.23
Block	4.18	1	.04*
Group:Block	103.16	1	<.001*

Table S5.

Wald test results for the final conflict responses, assessing the statistical significance of the fixed effect of the model with a 95% confidence interval. The significant p-values ($p < .05$) are marked with a *.

	Chi Square	Df	*p*
Group	3.59	1	.058
Block	0.19	1	.67
Group:Block	88.19	1	<.001*

H. Accuracy comparison in the current study (left panel) and in Franiatte et al.'s (2024) study (right panel)

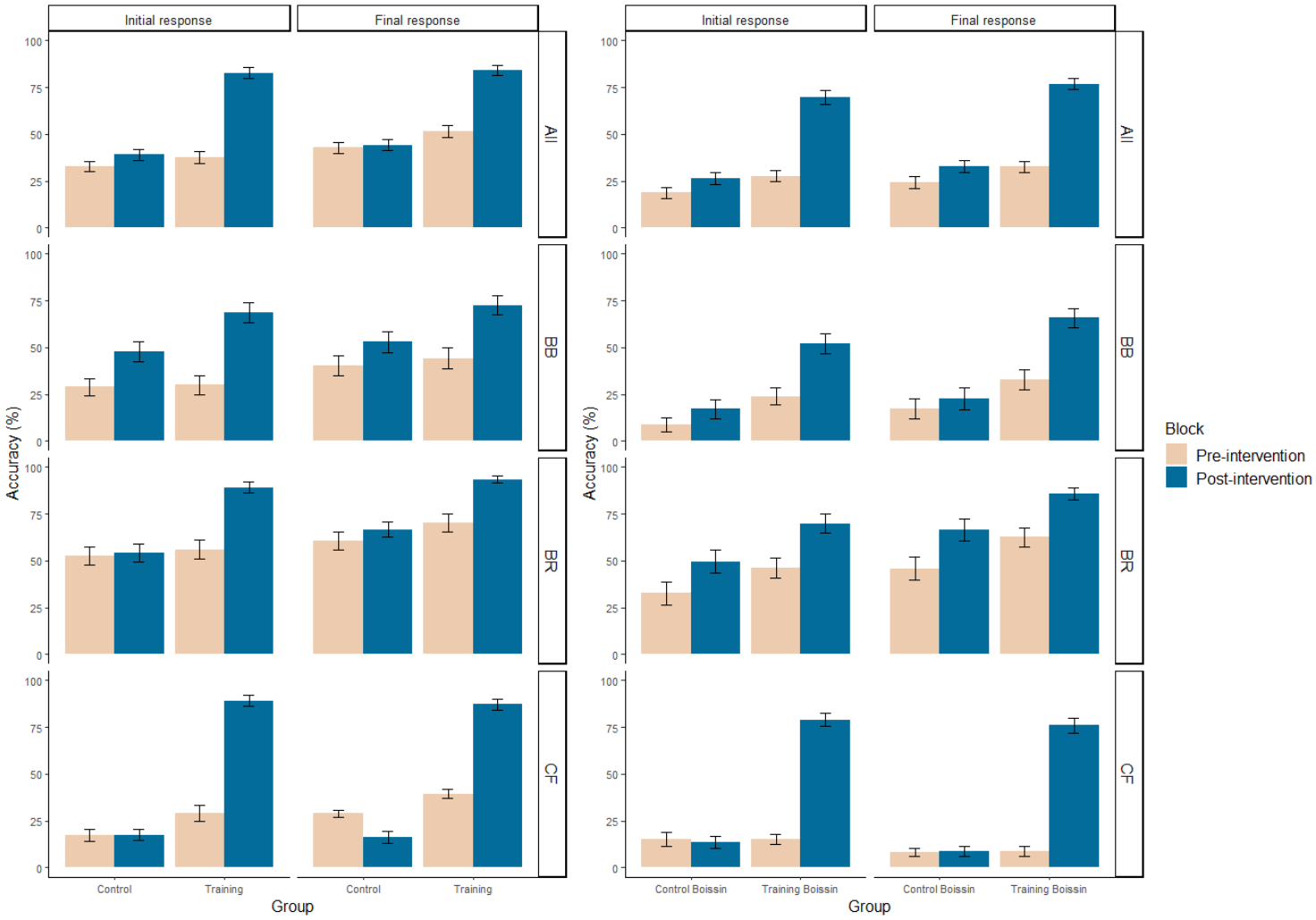


Figure S3. Comparison of mean accuracies (%) of correct initial and final responses on conflict problems for control and training groups, in the current study (left panel) and in Franiatte et al.'s (2024) study (right panel), for each task (BB, BR, CF), and combined (All). BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy, All = the composite mean across the three tasks.

Note. In Franiatte et al.'s study, the mean accuracies presented here corresponds to those of Session 1, Study 1.

I. Direction of change comparison with previous debiasing studies

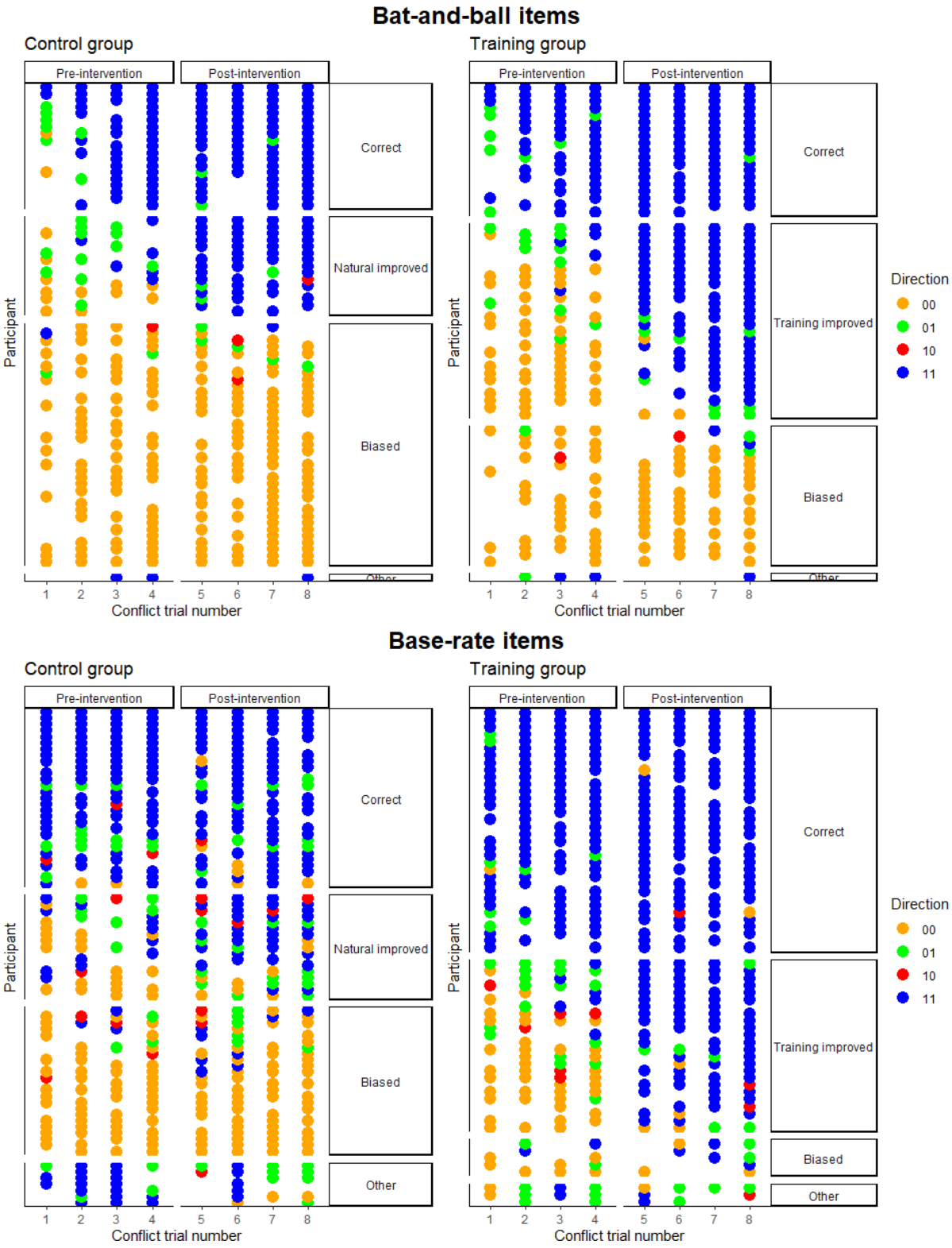
***Table S6*.**

Comparison between proportions (%) of each direction of change (i.e., '00' trials, '01' trials, '10' trials and '11' trials) for the conflict problems (SD), in each block (pre- and post-intervention) and each group (control and training), in Franiatte et al.'s (2024) study and in the current study.

Direction	Group	Block	Direction of change in Franiatte et al. (2024)	Direction of change in the current study
00	Control	Pre-intervention	74.7 (28.3)	51.1 (26.1)
00	Control	Post-intervention	63.8 (23.3)	51.0 (27.1)
00	Training	Pre-intervention	64.7 (29.5)	43.0 (27.1)
00	Training	Post-intervention	19.0 (26.6)	11.5 (18.8)
01	Control	Pre-intervention	10.3 (16.5)	15.7 (12.1)
01	Control	Post-intervention	10.2 (16.2)	9.8 (11.9)
01	Training	Pre-intervention	9.1 (12.5)	19.2 (14.3)
01	Training	Post-intervention	11.6 (17.8)	5.7 (11.8)
11	Control	Pre-intervention	11.6 (20.4)	27.5 (24.1)
11	Control	Post-intervention	21.8 (21.3)	33.7 (25.1)
11	Training	Pre-intervention	22.6 (27.0)	32.2 (26.6)
11	Training	Post-intervention	64.2 (33.1)	79.1 (27.0)
10	Control	Pre-intervention	3.4 (8.4)	5.6 (8.5)
10	Control	Post-intervention	4.1 (7.3)	5.5 (7.2)
10	Training	Pre-intervention	3.6 (8.2)	5.6 (7.5)
10	Training	Post-intervention	5.2 (8.8)	3.6 (6.3)

Note. In Franiatte et al.'s study, the mean accuracies presented here corresponds to those of Session 1, Study 1.

J. Individual level direction of change



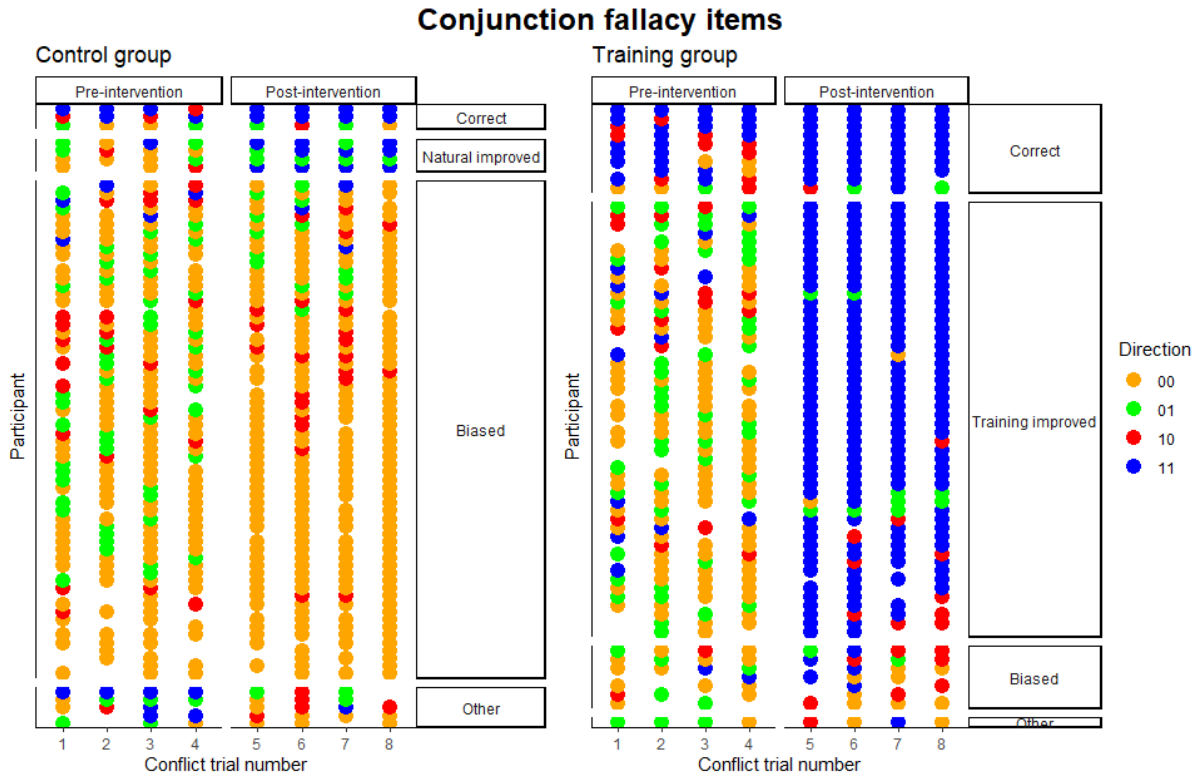


***Figure S4*.** Individual level direction of change (each row represents one participant). Due to the exclusion of missed deadline and load trials (see Trial Exclusion), not all participants contributed 24 analysable trials.