JOURNAL OF COGNITIVE PSYCHOLOGY
https://doi.org/10.1080/20445911.2023.2181734

No easy fix for belief bias during syllogistic reasoning?

Esther Boissin [a], Serge Caparos [b,c] and Wim De Neys [a]

[a]Université Paris Cité, LaPsyDÉ, CNRS, Paris, France; [b]Université Paris 8, DysCo Lab, Saint-Denis, France; [c]Institut Universitaire de France, Paris, France

ABSTRACT

Although erroneous intuitions often lead human thinking astray, recent studies suggest that single-shot interventions in which the underlying problem logic is clarified can easily remediate this bias. Because previous work typically focused on numerical problems, we tested here the generalizability to the infamous non-numerical belief bias during syllogistic reasoning. Unfortunately, results of 3 studies show that the effect is less clear. Although we succeeded in boosting performance for a minority of reasoners, reasoners who remained biased also tended to show a worse performance after training. We conclude that more work is needed to optimise the single-shot "easy fix" intervention approach to remediate belief bias during syllogistic reasoning.

ARTICLE HISTORY
Received 6 October 2022
Accepted 13 February 2023

KEYWORDS
Reasoning; heuristics & biases; de-biasing; intuition; syllogistic reasoning

Introduction

Human reasoning frequently relies on intuitions based on our subjective prior beliefs, which can sometimes offer us valid problem solutions but can also skew our judgment. For instance, while the Covid virus can affect any human being, irrespective of their ethnicity, thinking that the Chinese were somehow linked to the pandemic led some people to change their attitudes and behaviours, such as avoiding shaking hands with Asian individuals (Koller et al., 2021) or verbally and physically assaulting them (e.g. Gao & Liu, 2021). Similarly, the more stereotypically black a defendant is perceived to be, the more likely that person is to receive the death penalty (Eberhardt et al., 2006). This common human tendency to base judgments on personal beliefs and intuitions rather than on logical reasoning, biases performance in many classical reasoning tasks (Evans, 2003; Kahneman et al., 1982).

Cognitive scientists have long tried to remediate people's biased thinking and get them to reason correctly (e.g. Lilienfeld et al., 2009; Milkman et al., 2009; Nisbett, 1993). A number of recent studies have been especially successful in this respect (e.g. Boissin et al., 2021, 2022; Bourgeois-Gironde & Van Der Henst, 2009; Claidière et al., 2017; Hoover & Healy, 2017, 2019; Morewedge et al., 2015; Purcell et al., 2020; Trouche et al., 2014). These "debiasing" studies have shown that a short single-shot explanation about the intuitive bias and the correct solution strategy often helps reasoners to solve structurally similar problems afterwards.

However, as of today debias effects have been observed only for a handful of reasoning problems, such as bat-and-ball problems (e.g. Boissin et al., 2021; Bourgeois-Gironde & Van Der Henst, 2009; Hoover & Healy, 2017; Purcell et al., 2020), CRT or CRT-2 items (Isler et al., 2020; Trouche et al., 2014), base-rate problems (Boissin et al., 2022) or conjunction fallacy problems (Boissin et al., 2022; Claidière et al., 2017). These problems are typically based on mathematical or probabilistic components. Given the potential implications of debiasing human thinking, further validation is necessary with non-numerical problems. The first aim of the present study was to examine the robustness and generalizability of the debiasing effect to logical, non-numerical problems. In the literature, one of the most tested examples of non-numerical skewed reasoning is the belief-bias effect in syllogistic reasoning.

Belief bias refers to the intuitive tendency to judge the validity of a syllogism on the basis of

CONTACT Esther Boissin boissinesther@gmail.com LaPsyDE (CNRS & Université Paris Cité), Sorbonne – Labo A. Binet, 46, rue Saint Jacques, 75005 Paris, France

Supplemental data for this article can be accessed online at https://doi.org/10.1080/20445911.2023.2181734.

the believability of its conclusion—or empirical validity—rather than on its logical validity (Oakhill et al., 1989). This heuristic is problematic because the believability of a syllogism's conclusion is independent of its logical validity. Consider the following example: "All things that are smoked are bad for your health. Cigarettes are bad for your health. Therefore, cigarettes are smoked". Although the conclusion in the example is logically invalid and should be rejected, intuitively many people nevertheless tend to accept it because it fits with their prior beliefs.

Note that previous debias work on belief bias has not always been successful. Most prominently, in a series of experiments in the 1990s, Evans, Newstead, and colleagues (Evans et al., 1994; Newstead et al., 1992) observed that short instructions and clarifications of the syllogistic problem structure led to mixed results (i.e. small, no, or non-replicated effects). Hence, more work is needed to test whether there is an easy fix (i.e. single shot training) to belief bias.

The present study also aimed to explore the nature of possible belief bias debiasing intervention effects. A key question is whether the training intervention primarily affects people's intuitive or deliberate thinking (or, in the popular dual-process terms, their fast "System 1" or slow "System 2", e.g. Kahneman, 2011). The common assumption is that after training, participants are more likely to engage deliberative processes and correct their intuitively generated intuitive responses (e.g. Evans, 2019; Lilienfeld et al., 2009; Milkman et al., 2009). However, it is also possible that once reasoners grasp the correct solution, they will no longer generate an incorrect response and intuit correctly right away, with no need for a corrective "System 2" deliberation process.

Support in favour of the "trained intuitor" viewpoint is currently accumulating (Boissin et al., 2021, 2022; Purcell et al., 2020). To differentiate intuitive responses from deliberate ones, studies have used a so-called two-response paradigm (Thompson et al., 2011). 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 initial intuitive hunch that comes to mind. Next, they can take all the time they need to reflect on the problem and give a final deliberate response. To make maximally sure that the initial response is generated intuitively, the response needs to be given under time pressure and/or cognitive load (i.e. resources that are critical to engage in deliberation, Bago & De Neys, 2017). This paradigm allows to measure the training impact both on people's intuitive reasoning performance (i.e. initial response accuracy) and on deliberate reasoning performance (i.e. final response accuracy). Overall, the two-response results showed that debias training typically already boosted correct responses in the initial, intuitive response stage (Boissin et al., 2021, 2022).

These de-bias findings may have important theoretical and applied implications (Boissin et al., 2021, 2022). If a short, single shot intervention manages to remediate people's intuitive thinking, this implies that the bias is less profound than it has long been assumed (Kahneman, 2011). Likewise, it presents us with a straightforward pedagogical tool to optimise reasoning and avoid the perils of biased inferencing on people's thinking altogether. However, before drawing strong conclusions it is critical to test the generality of the findings and establish the possible limitations of the approach.

In the present work, we aimed to investigate whether a training intervention can help participants produce correct intuitions for syllogistic problems to test whether the findings extend beyond non-numerical reasoning problems. In Studies 1 and 2, we compared participants' reasoning performance before and after receiving a short training session, using a two-response paradigm. We contrasted their performance to that of participants who received no training (the control group). In Study 3, participants were re-tested after a two-months delay in order to examine whether the training effect sustained over time.

Study 1

Method

Pre-registration and data availability

The design and research questions of Studies 1, 2 and 3 were preregistered on the AsPredicted website (https://aspredicted.org) and stored on the Open Science Framework (https://osf.io/rzm92/) where all data and material can be accessed. No specific statistical analyses were preregistered.

Participants

Participants were recruited online, using the Prolific Academic website (http://www.prolific.ac). Participants had to be native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom to take part. In total, 99 individuals participated (69 females, M =

35.4 years, SEM = 1.4), 49 participants were randomly assigned to the training group and 50 to the control group. In total, 43 participants had secondary school as their highest level of education, and 56 reported a university degree. We compensated participants for their time at the rate of £5 per hour.

Materials

The current study consisted of three blocks presented in the following order: A pre-intervention, an intervention, and a post-intervention. In total, each participant had to solve 24 problems, namely, four conflict, four no-conflict, two neutral and two transfer problems during the pre-intervention, and again the same number of problems during the post-intervention. All the problems are presented in Supplementary Material Section A.

Syllogism. Syllogisms were taken from Brisson et al. (2018) and were categorical versions of two conditional inferences. The first was equivalent to the valid *Modus Ponens* (If P then Q, P is true, therefore Q is true) and the second was equivalent to the invalid *Affirmation of the Consequent* (If P then Q, Q is true, therefore P is true).

Each problem included a major premise (e.g. "All mammals can walk"), a minor premise (e.g. "Whales are mammals") and a conclusion (e.g. "Whales can walk"). Participants were instructed that they had to accept that the premises were true. The task's goal was to evaluate whether the conclusion follows logically from the premises. The following format was used for (1) the valid problem whose conclusion follows logically and for (2) the invalid problem whose conclusion does not follow logically:

(1) *All mammals can walk.*
Whales are mammals.
Whales can walk.
Does the conclusion follow logically?
- *Yes*
- *No*

(2) *All mammals can walk.*
Dogs can walk.
Dogs are mammals.
Does the conclusion follow logically?
- *Yes*
- *No*

For half of the items, believability and validity of the conclusion conflicted (conflict items) while for the other half, conclusion validity was in accordance with its believability (no-conflict items). The followings are examples of a no-conflict (3) valid and (4) invalid problem:

(3) *All mammals can walk.*
Cats are mammals.
Cats can walk.
Does the conclusion follow logically?
- *Yes*
- *No*

(4) *All mammals can walk.*
Birds can walk.
Birds are mammals.
Does the conclusion follow logically?
- *Yes*
- *No*

These no-conflict control 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).

Pre- and post-intervention were each composed of four conflict items (two problems with an unbelievable—valid conclusion, and two problems with a believable—invalid conclusion) and four no-conflict items (two problems with a believable—valid conclusion, and two problems with an unbelievable—invalid conclusion). In each block, two sets of items were used for counterbalancing purposes. Valid believable and unbelievable conclusions in one set were invalid in the other set and vice-versa. Items were presented in a randomised order.

Justification. After the last problem of the post-intervention, which was always a believable-invalid conflict problem, participants were asked to type in a justification for their final response (see Supplementary Material Section B for further details). Results indicated that the majority of correct responses was indeed correctly justified (training group: 10 correct justification out of 19 correct responses, control group: 13 correct justifications out of 19 correct responses, see Supplementary Material Section B). Note that the justification was untimed and retrospective. It was collected for exploratory purposes and does not allow us to draw any conclusion with respect to the intuitive or deliberate nature of participants' processing.

Neutral problems. We also presented four neutral problems. Those items used abstract content with the

same logical structure as the other syllogistic items (e.g. "All WWW are YYY…"). They were either in valid or invalid form. The neutral items are traditionally used to track people's knowledge of the underlying logical principles or "mindware" (Stanovich, 2011). When people are allowed to deliberate, reasoners have little trouble solving them (De Neys & Glumicic, 2008; De Neys et al., 2008; Frey et al., 2017, 2018). The neutral problems allowed us to explore whether a potential learning effect on conflict syllogisms in which the reasoner needs to discard a conflicting prior subjective belief, leads to a more general performance boost on other untrained syllogistic problems that do not contain information involving subjective knowledge.

Transfer problems. In addition to the syllogisms, we presented another type of reasoning problem to test whether the "syllogism" training effect could transfer to untrained problems. In total, we used two *Modus Tollens* taken from Brisson et al. (2018) and two conjunction-fallacy problems from Frey et al. (2018). We presented one *Modus Tollens* and one conjunction-fallacy problem at the end of the pre-intervention and again at the end of the post-intervention.

The *Modus Tollens* have negated forms of the valid problem, composed of unbelievable conclusions which conflicted with the validity. Here is an example:

All things with four legs are dangerous.
Poodles are not dangerous.
Poodles do not have four legs.
Does the conclusion follow logically?
- *Yes*
- *No*

For each of the two conjunction problems, participants were given a short personality description of an individual and were asked to indicate which of the two statements was most probable. One statement always consisted of a conjunction of two characteristics (one characteristic that was likely given the description, i.e. a stereotypical association, and one that was unlikely). The other statement contained only the unlikely characteristic. The following illustrates the structure of the conjunction problem:

Jake is 20.
He grew up in a poor family in a neglected neighbourhood.
He is quite violent and already served a short sentence in prison.
Which statement is most likely?
- *Jake plays the violin*
- *Jake plays the violin and is jobless*

Given that the conjunction of two events cannot be more likely than each of the constituent events (formally: $p(A\&B) \leq p(A)$) the correct response was the non-conjunctive statement.

Intervention. During the intervention, the participants tried to solve four syllogistic problems. Two syllogisms were believable-invalid conflict items and two were unbelievable-valid conflict items with the same structure as the pre- and post-intervention problems. Participants in the training group were explained the correct solution after having given their response to each problem. Participants in the control group received no such explanation. The following illustrates the explanation about one believable-invalid item (e.g. "All things that have a motor need oil. Cars need oil. Cars have motors".):

> The correct answer to the previous problem is that the conclusion does not follow logically. Many people think it does, but this answer is wrong.
>
> Most people base their answer solely on the content of the conclusion. They accept conclusions that are believable (e.g. "Cars have motors") and reject conclusions that are unbelievable (e.g. "Whales can walk"). However, in order to assess whether a conclusion is logically correct, you need to focus solely on the underlying logical structure.
>
> An argument of the structure "All X are Y" (e.g. "All things that have a motor need oil") implies that everything that is said to be an X (e.g. "Teslas have motors") is always Y (e.g. "Teslas need oil") whether or not this is actually believable. However, the reverse does not hold. If I say that "All X are Y" (e.g. "All things that have a motor need oil") it does not follow that everything that is a Y (e.g. "Cars need oil") is also an X (e.g. "Cars have motors"), even though the conclusion might sound believable.
>
> In sum, logically speaking the statement "All X are Y" implies that whenever you have the first part (something is X), the second part follows logically. However, it does not imply that whenever you have the second part (something is Y), the first part also follows.

Participants also always received explanations about the correct solution of unbelievable-valid

items. For example, for the item: "All books are made of paper. E-books are books. E-books are made of paper", the following explanation was displayed:

> The correct answer to the previous problem is that the conclusion follows logically. Many people think it does not, but this answer is wrong.
>
> Most people base their answer solely on the content of the conclusion. They accept conclusions that are believable (e.g. "Strawberries are fruits") and reject conclusions that are unbelievable (e.g. "E-books are made of paper"). However, in order to assess whether a conclusion is logically correct, you need to focus solely on the underlying logical structure.
>
> An argument of the structure "All X are Y" (e.g. "All books are made of paper") implies that everything that is said to be an X (e.g. "E-books are books") is always Y (e.g. "E-books are made of paper") whether or not this is actually believable. However, the reverse does not hold. If I say that "All X are Y" (e.g. "All books are made of paper") it does not follow that everything that is a Y (e.g. "Encyclopaedia are made of paper") is also an X (e.g. "Encyclopaedia are books"), even though the conclusion might sound believable.
>
> In sum, logically speaking the statement "All X are Y" implies that whenever you have the first part (something is X), the second part follows logically. However, it does not imply that whenever you have the second part (something is Y), the first part also follows.

The explanations were based on the same general principles that were adopted by Boissin et al. (2021, 2022): The explanations 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 incorrect response. No personal performance feedback (e.g. "Your answer was wrong") was given in order to avoid promoting feelings of judgment (Trouche et al., 2014). The intervention block always began by a believable-invalid item and its explanation, followed by an unbelievable-valid item and its explanation. Participants moved on to the following screen by clicking on the "Next" button.

Two-response format. For both the pre- and post-intervention, participants responded to each problem using a two-response procedure, where they first provided a "fast" answer, directly followed by a second "slow" answer (Thompson et al., 2011). This method allowed us to capture both an initial "intuitive" response, and then a final "deliberate" one. To minimise 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 (see Bago & De Neys, 2017, 2019; Raoelison & De Neys, 2019). The load task was based on the dot memorisation 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; Verschueren et al., 2004). Participants had to memorise a complex visual pattern (i.e. 4 crosses in a 3×3 grid) presented briefly before each reasoning problem. After their initial (intuitive) response to the problem, participants were shown four different patterns (i.e. with different matrices of crosses) and had to identify the one that they had memorised (see Bago & De Neys, 2017, for more details).

For all syllogistic problems, a time limit of 3 seconds was used for the initial response, based on previous pretesting that indicated it amounted to the time needed to read the preambles, move the mouse, and select an answer (Bago & De Neys, 2017; Raoelison et al., 2020). For the lengthier transfer conjunction problem, the time limit was set to 6 seconds. Time limit and cognitive load were applied only for the initial response, and not for the final one (see below).

Procedure

The experiment was run online using the Qualtrics platform. Participants were instructed that the experiment would take 15 minutes and that it demanded their full attention. A general description of the task was presented in which participants were instructed that they would read reasoning problems, for which they would have to provide two consecutive responses. They were told 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 (see Bago & De Neys, 2017, for literal instructions). In order to familiarise themselves with the two-response procedure, they first solved two unrelated practice problems. Next, they familiarised themselves with the cognitive load procedure 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 syllogism trial. We adopted the presentation format of Bago and De Neys (2017). All trials started with the presentation of a fixation cross for 2000ms, followed by the first sentence (i.e. the major premise), e.g. "All mammals can walk", for 2000ms, and subsequently, by the visual matrix for the cognitive-load task for 2000ms. Afterwards, the second sentence (i.e. minor premise), e.g. "Whales are mammals", was presented under the first premise for 2000ms, followed by the full problem which featured the conclusion with the question "Does the conclusion follow logically?" And the two response options (yes/no). At this point participants had 3000 ms to choose a response. After 2000 ms the background of the screen turned yellow 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 how confident they were about their response (from 0%, absolutely not confident, to 100%, absolutely confident). Then, they were presented with four visual matrices and had to choose the one that they had previously memorised. They received feedback as to whether their memory response was correct. If the answer was not correct, they were reminded that it was important to perform well on the memory task on subsequent trials. Finally, the same reasoning problem was presented again, and participants were asked to provide a final deliberate answer (with no time limit) and, once again, to indicate their confidence level.

Note that, given the different nature of the transfer conjunction problems, we adopted slightly different timings than for the initial response of the syllogistic problems. To begin with, the problems appeared in three parts. The first part of the conjunction fallacies remained on screen for 2000 ms (e.g. "Jake is 20. He grew up in a poor family in a neglected neighbourhood"). Then, the visual matrix appeared for 2000ms and next, the entire problem was displayed (e.g. "Jake is 20. He grew up in a poor family in a neglected neighbourhood. He is quite violent and already served a short sentence in prison") and remained on screen for 2000ms. Afterwards, the question (e.g. "Which statement is most likely?") and the two responses options (e.g. "Jake plays the violin" or "Jake plays the violin and is jobless") were shown for another 6000 ms. After 4000 ms the background turned yellow to warn participants for the deadline.

At the end of the study, participants in the control group were also presented with the explanations about how to solve the syllogistic problems, and all participants were asked to complete their demographic information.

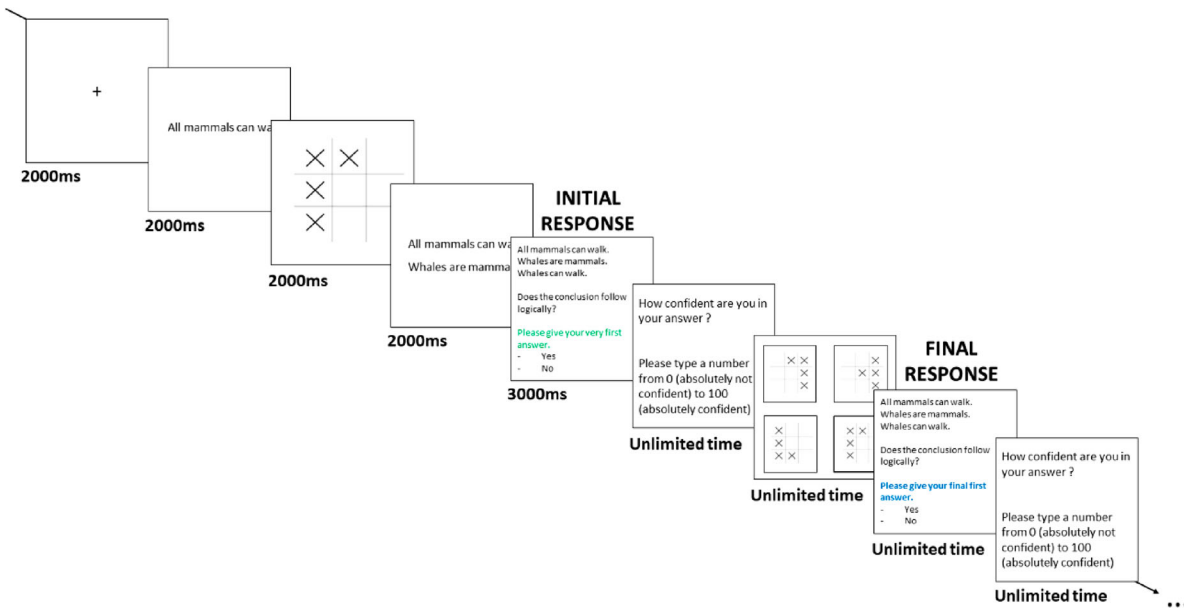

Figure 1. Time course of a complete two-response syllogistic item.

Trial exclusion

We discarded trials in which participants failed to provide their initial answer before the deadline (7.7% of all trials) or failed to pick the correct matrix in the load task (14.0% of the remaining trials), and we analysed the remaining 79.3% of all trials. On average, each participant contributed 19.7 (SEM = 0.3) trials out of 24.

Results and discussion

Syllogism response accuracy. For each participant, we calculated the average proportion of correct initial and final responses, for the conflict and no-conflict problems, in each of the two blocks (pre- and post-intervention). We analysed the data using mixed-design ANOVAs on initial and final accuracies with Block (pre- VS post-intervention) as a within-subjects factor and Group (training VS control) as a between-subjects factor.

Figure 2 shows the results. As the figure shows, on the critical conflict problems there is a slight tendency towards an increased post-intervention performance in the training group, both for the initial (+7.3 points) and final (+11.0 points) trials. Note however, that this effect was much more restricted than on similar previous debias studies with numerical reasoning problems (e.g. bat-and ball, base-rate, and conjunction fallacy items in Boissin et al., 2021, 2022) on which training led to an accuracy increase of up to 58.3 points for final responses and 48.3 points for initial responses. The ANOVA indicated that the central Block x Group interaction reached significance for the final trials, $F(1,95) = 10.79$, $p = .001$, $\eta^2g = .0$, while it only showed a marginal effect for initial trials, $F(1,94) = 3.17$, $p = .08$, $\eta^2g = .01$.

In and by itself a small training effect on conflict problems might still be meaningful. However, as Figure 2 (bottom panel) shows, the problem in the current study was that the small training increase on the conflict problems was accompanied by an approximately similar post-intervention decrease on the control no-conflict problems for the training group, also both for initial (−9.7 points) and final (−8.9 points) trials. The no-conflict problems are expected to be easily solvable if one pays sufficient attention and, in theory, training should

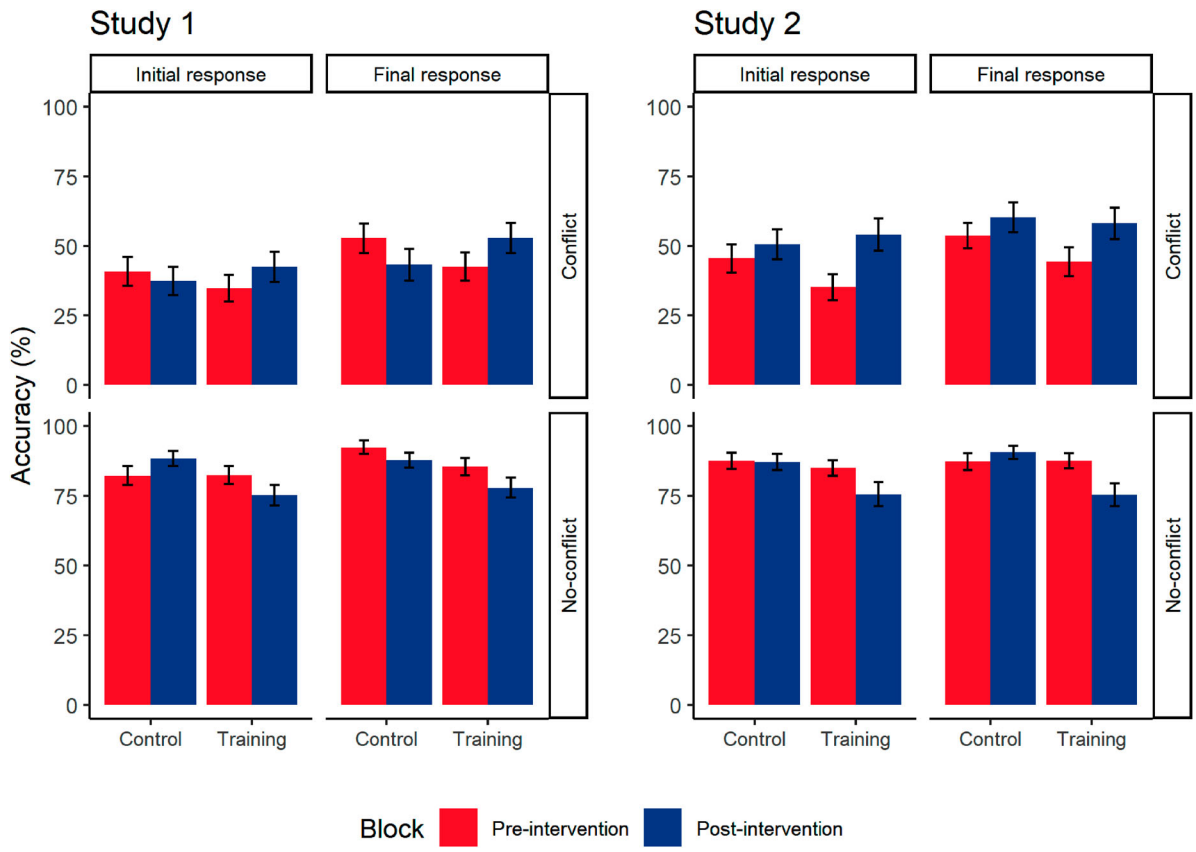

Figure 2. Average initial and final response accuracies on conflict and no-conflict problems in Study 1 and 2, for each group (i.e. Control VS Training), before and after the intervention. Error bars represent standard errors of the mean (SEM).

have no impact on performance. An ANOVA on no-conflict accuracy pointed to a significant Block x Group interaction for the initial, $F(1,91) = 6.89$, $p = .01$, $\eta^2g = .03$ trials, although not for the final, $F(1,94) = 0.80$, $p = .37$, $\eta^2g = .00$, ones.

In sum, the data indicate that the post-intervention performance improvement observed with conflict problems went together with a decrease in performance for no-conflict problems. This finding suggests that the training may have simply cued participants into using a heuristic that led them to consider that a believable answer is always incorrect, and vice versa for unbelievable answers. This would then cause participants to reject believable conclusions and accept unbelievable conclusions for all problems, regardless of the logical validity of the conclusion. While for conflict problems this strategy would result in correct responses, it would result in incorrect responses for no-conflict control problems.

The pre- and post-interventions were composed of a mix of two types of logical structures, i.e. the valid and invalid syllogisms, presented in a random order. We investigated if the deleterious effect of training on no-conflict performance occurred equally for the two types of logical structure. In order to do so, we contrasted no-conflict performance before and after the training across the two types of items. An identical response pattern was detected for both the invalid and valid syllogisms (See Supplementary Material, Section C) meaning that the drop in no-conflict accuracy after the intervention was not explained by a decrease in just one given item type. This suggests that the deleterious effect was driven by a confusion which occurred due to the mix of two different items, rather than the effect of one specific item type.

Individual level directions of change. To gain some deeper insight into how people changed (or did not change) their response after deliberation, we performed a direction of change analysis (Bago & De Neys, 2017, 2019). On each trial, participants could give either a correct ("1") or incorrect ("0") response, at each of the two response stages (i.e. initial and final). Hence, this can result in four different types of response patterns on any single trial ("00" pattern, incorrect response at both stages; "11" pattern, correct response at both stages; "01" pattern, initial incorrect and final correct response; "10" pattern, initial correct and final incorrect response).

For each participant, on each conflict trial, we coded the direction of change from start to end of the experiment. This allowed us to observe, at a higher level of detail, how the intervention influenced participants' response patterns.

By and large, as in previous studies (e.g. Boissin et al., 2021, 2022; Raoelison & De Neys, 2019), Figure 3 suggests that we can classify the participants into three main categories. First, 41.7% of the participants did not benefit from the training intervention since they gave a majority of incorrect (biased) responses (i.e. "00" patterns) both before and after the intervention. These participants were classified as "biased" respondents. Second, some participants gave a majority of correct initial and/or final responses (i.e. "01" or "11" patterns) from start to finish and did not require any training intervention to respond correctly to the syllogistic problems. They represent 18.8% of the participants and were labelled as "correct" respondents in Figure 3. Third, some participants improved their performance after the intervention and were labelled as "improved" respondents. These were participants who showed a post-intervention increase in "01" patterns (at the expense of "00" patterns), or an increase in "11" patterns (at the expense of either "00" or "01" patterns). Overall, improved respondents represented a small proportion of the training group, namely, 22.9%. Note that, in the training group, some participants showed an inconsistent response pattern and could not be classified based on our criteria. They were put in an "other" group and represent 16.7% of all training group.

Note that in the control group about 4.1% of reasoners showed a natural improvement, in the absence of training, and started giving correct responses after the control ("no-explanation") intervention block. These participants were labelled as "natural improved". In and by itself, this natural-improved group (4.1% of reasoners) was considerably smaller than the improved group in the training condition (22.9% of reasoners).

Accuracy across training respondent categories. Study 1 showed that, in the training group, the short explanation given during the intervention improved overall performance on conflict problems. However, this positive effect was accompanied by a negative effect on no-conflict (control) problems, for which performance should have remained high throughout (Bago & De Neys, 2019; Boissin et al., 2021, 2022; Brisson

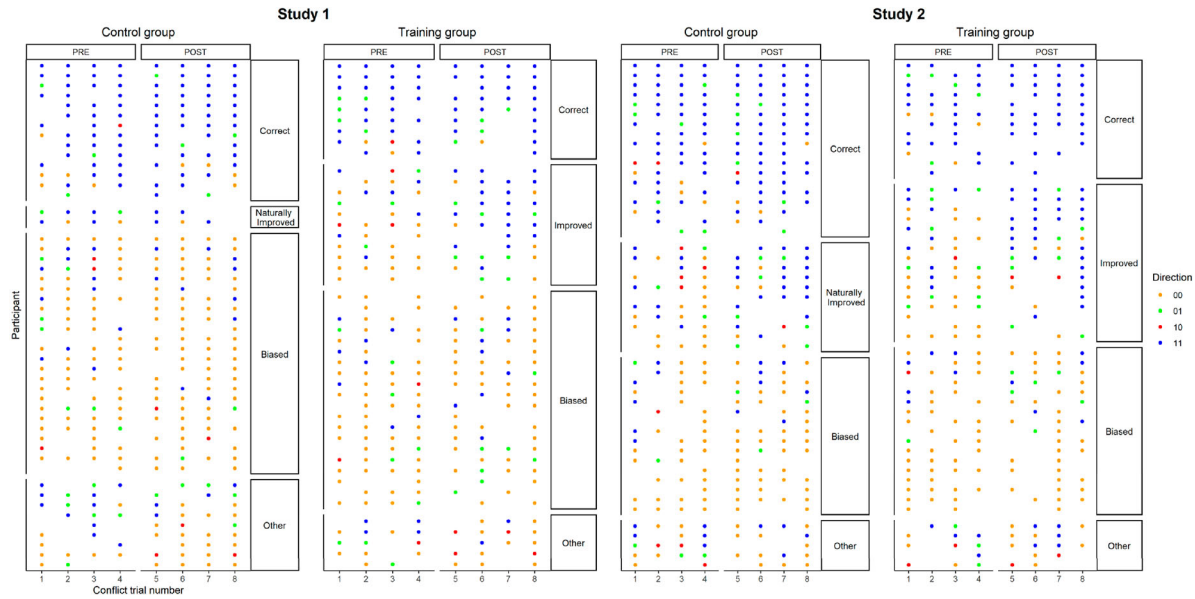

Figure 3. Individual level direction of change of Study 1 and Study 2. Each row represents one participant. Each point represents one response from start to finish of the study (left to right). Participants are classified according to the training effect (i.e. Correct reasoners give a majority of "11" trials throughout the study, Improved reasoners give a majority of "00" trials or "01" trials before the intervention and a majority of "01" trials or "11" trials after the intervention, Biased reasoners give a majority of "00" trials before and after the intervention, and "Other" reasoners show an inconsistent response pattern). Due to the discarding of missed deadline and load trials (see Trial Exclusion), not all participants contributed 8 analysable trials.

et al., 2018; Raoelison et al., 2020, 2021). To verify that this negative effect was a result of the intervention, we examined performance on no-conflict problems across the different types of respondents in the training group (i.e. biased VS improved VS correct).

Figure 4 shows that correct respondents (i.e. those who did not need any training to solve conflict syllogisms correctly) performed similarly before and after training on no-conflict problems. We observed the same pattern in biased respondents (i.e. those who did not benefit from training and provided incorrect responses to conflict items throughout the experiment). However, improved reasoners (i.e. those whose accuracy on conflict problems increased after training) showed a post-intervention drop in performance with no-conflict items. As suggested, this indicates that our slight performance boost on the conflict problems presumably results from the application of a heuristic that led reasoners to infer that a believable answer is always incorrect, and vice versa for unbelievable answers.

Note, as one reviewer suggested, to quantify how many individuals showed this pattern we simply tallied what percentage of reasoners in the improved group showed a no-conflict accuracy decrease that was smaller than their conflict problem accuracy increase. This can be used as a proxy to disentangle the proportion of improved reasoners who genuinely benefitted from training or applied the erroneous believability heuristic. Results showed that 64% of the improved group genuinely benefit from the training (either at the intuitive and/or deliberate stage).

Hence, although we managed to genuinely improve the performance of a small subgroup of reasoners, we did not manage to boost the application of the underlying logical principle per se overall.

Conflict detection. Previous studies have shown that despite giving an incorrect response, reasoners sometimes detect their error or the presence of a response conflict (e.g. De Neys et al., 2013; Frey et al., 2017). In this study, we explored whether the training intervention affected biased reasoners' ability to detect conflict in syllogisms. That is, although the training might not have succeeded in getting biased people to reason accurately, it might have helped them to better detect that their answer was incorrect. We used the conflict-detection index introduced in the study of De

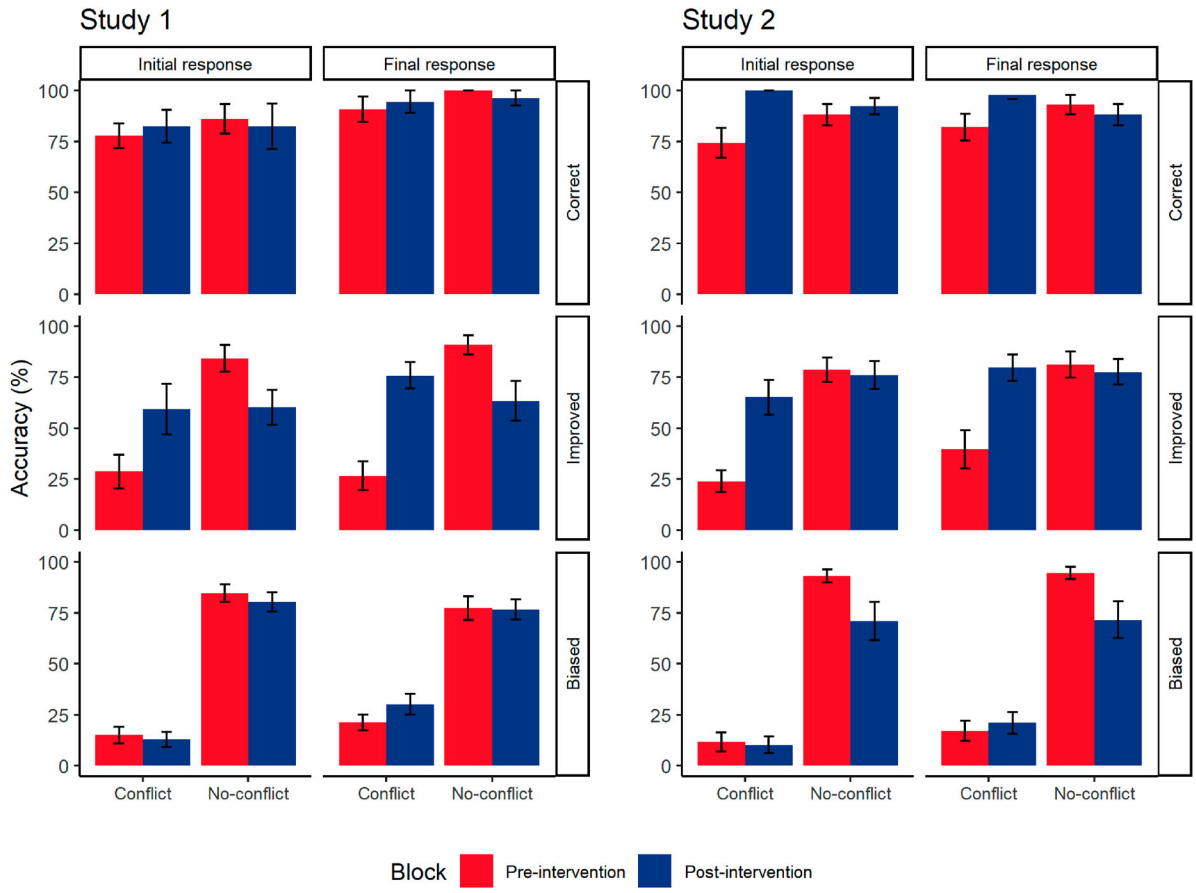

Figure 4. Average initial and final response accuracies on conflict and no-conflict problems in Study 1 and 2 for each type of training respondents (i.e. Correct, Improved and Biased respondents), before and after the intervention. Error bars represent standard errors of the mean (SEM).

Neys et al. (2011), which contrasts confidence[1] ratings for no-conflict trials that yielded a correct response to confidence ratings for conflict trials that yielded an incorrect response. We compared the conflict-detection index before and after the intervention, in both the training and control groups. A higher difference value implies a larger confidence decrease when solving conflict items, which is believed to reflect a more pronounced conflict experience (Bago & De Neys, 2019; Pennycook et al., 2015).

Table 1 indicates that neither for initial, nor final responses, the intervention affected conflict detection. The Group x Block interaction failed to reach significance, for both final responses, $F(1,61) = 0.05$, $p = .83$, $\eta^2g = .00$, and initial ones, $F(1,69) = 0.12$, $p = .73$, $\eta^2g = .00$.

Predictive conflict detection. As suggested by a reviewer, we also investigated whether individual differences in the ability to detect conflict (before the intervention) was predictive of the success of the training intervention. That is, we tested whether reasoners who became correct respondents after the training intervention (i.e. improved respondents in our individual level classification) showed better conflict detection (i.e. stronger response doubt when giving incorrect responses on conflict problems) before the training compared to reasoners who did not improve after training (i.e. biased respondents). We again used the difference in confidence ratings for incorrect conflict problem responses and correct no-conflict control problem responses as our index of conflict detection. The higher the conflict detection index was, the more a participant doubted their incorrect response (i.e. the more they detect their error).

For initial responses, we observed a small trend indicating better conflict detection for improved

[1]Since it has been shown that the initial response latency is not a reliable measure for conflict detection (Bago & De Neys, 2017), we will only present the conflict detection associated with the confidence rates.

Table 1. Conflict detection (i.e, percentage of mean difference in confidence ratings (SEM) between conflict and no-conflict trials) results in Study 1 and Study 2.

	Group	Initial response		Final response	
		Pre-intervention	Post-intervention	Pre-intervention	Post-intervention
Study 1	Control	4.22% (4.02)	6.73% (2.82)	2.56% (2.77)	1.72% (1.99)
	training	0.06% (2.27)	4.88% (3.36)	−0.06% (2.66)	0.60% (2.38)
Study 2	Control	3.43% (3.55)	0.40% (2.63)	−5.96% (3.06)	−2.19% (2.24)
	training	5.70% (3.28)	4.37% (3.27)	4.56% (2.36)	−1.40% (2.36)

respondents (M = 3.7%, SEM = 2.9) compared to biased respondents (M = −3.8%, SEM = 2.5), t(27) = 1.76, $p = .09$, d = .68. There was no difference for the final responses (M improved = 1.2%, SEM = 2.6; M biased = −1.5%, SEM = 4.1; t(28) = 0.44, $p = .67$, d = .16). Note that, for both initial and final responses, reasoners from the biased group did not show a nominal detection effect (i.e. the conflict detection index was negative), showing that these participants did not doubt their incorrect conflict responses whatsoever.

Neutral problem accuracy. We also tested whether the training led to a performance increase with neutral problems, which assessed logical knowledge (or "mindware") in the absence of any conflict. There was no clear sign of a training effect on neutral problems (All $Fs < 1.30$, $ps > .20$, see Supplementary Material, Section D). In sum, the syllogistic training intervention did not boost reasoners performance on neutral problems.

Transfer problem accuracy. Finally, we explored whether the training intervention possibly improved performance for two types of non-trained reasoning problems, one that had a syllogistic format (i.e. Modus Tollens), and one that did not (i.e. conjunction fallacy problems). There was no effect of the training intervention, for either one of the response stages (All $Fs < 0.98$, $ps > .16$, see Supplementary Material, Section D). Hence, the results suggest that the increased conflict problem performance was specific to the trained syllogisms and did not lead to an increase in performance on untrained types of problems.

Study 2

The goal of Study 1 was to explore whether a training intervention decreased the influence of belief bias in a syllogistic reasoning task. While the training slightly improved performance on conflict problems, it also had an unexpected negative impact on performance with easily solvable no-conflict problems. The specific nature of the syllogistic training we adopted may account for the somewhat deleterious effect. In contrast with previous training studies with numerical problems (e.g. Boissin et al., 2021, 2022), in Study 1 participants had to assimilate two different logical structures, respectively the valid and invalid logical rules. These were introduced together in our explanatory intervention texts. It is possible that this led some participants astray, causing them to evaluate believable conclusions as always invalid and unbelievable ones as always valid, regardless of their logical structure. Consequently, the impact of training on syllogisms remains unclear.

The goal of Study 2 was to assess the impact of training with an optimised training design. In Study 2, the intervention consisted in the presentation of two consecutive mini training blocks in which each logical structure was explained and practiced in isolation with revised material. One training block focused on valid problems, while the other focused on invalid problems. We speculated that this might minimise possible logical structure interference.

Method

Participants

Participants were recruited online, using the Prolific Academic website (http://www.prolific.ac). Participants had to be native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom to take part. In total, 102 individuals participated (67 females and 1 non-gender, M = 35.8 years, SEM = 1.4), 51 participants were randomly assigned to the training group and 51 to the control group. Thirty-nine participants had secondary school as their highest level of education, and 63 reported a university degree. We compensated participants for their time at the rate of £7 per hour.

Materials and procedure

The pre- and post-intervention blocks of Study 2 used the same material and two-response procedure as that used in Study 1. For the intervention block, first, we simplified the explanations

provided to the participant during the training and, second, we used material for which conclusions were pre-tested in a pilot study. Thirty-four participants (18 females, M = 32.4 years, SEM = 2.3) were asked to rate the believability of each conclusion on a scale from 0 to 10 (0 being totally unbelievable and 10 being totally believable). The conclusions from the conflict and no-conflict problems were presented as separate conclusion to be evaluated. We selected the 16 items for which the mean believability was highest (closest to 10) or lowest (closest to 0).

In contrast with Study 1, in Study 2 participants were not asked to solve neutral or transfer problems, nor were they asked to justify their last response. In total, participants had to respond to 32 problems: 8 during the pre-intervention (4 conflict and 4 no-conflict problems) and the same number during the post-intervention. Participants also responded to 8 problems during each mini training session (2 conflict and 2 no-conflict before each explanation was given and again the same number after each explanation was given). All the problems are presented in Supplementary Material Section A.

Intervention. The intervention was split into two separate consecutive mini training sessions: One for the valid and one for the invalid problems. Each training session was split into three parts in the following order: A pre-explanation, an explanation and a post-explanation. Participants performed both mini trainings in a random order. Both pre- and post-explanation parts consisted of two conflict problems (in the case of valid problems, these problems had an unbelievable conclusion, while in the case of invalid problems, they had a believable conclusion) and two no-conflict problems (for valid problems, the conclusion was believable while for invalid problems, the conclusion was unbelievable). Pre- and post-explanation problems were presented using the two-response procedure. The explanation part consisted of presenting two problems in their conflict form followed by an explanation. Note that the second "explanation" part of the training session consisted solely in the presentation of two problems, with no explanation, in the control group.

The following illustrates an explanation about one believable-invalid item (e.g. "All reptiles are cold-blooded. Snakes are cold-blooded. Snakes are reptiles".):

> The correct answer to the previous problem is that the conclusion does not follow logically.
>
> Many people think that the conclusion follows logically, but this answer is wrong. Here is the explanation:
>
> All the problems you have just solved are made up of a logical structure and knowledge to which the conclusion refers. One way to conceptualize the logical structure of the problem you just answered above is to consider that:
>
> All A's are B's.
>
> All C's are B's.
>
> So all C's are A's.
>
> This logical structure always leads to the assumption that the conclusion does not follow logically.
>
> Most people only use their knowledge to decide whether the conclusion follows logically. Thus, they can answer that the conclusion follows logically if the conclusion is credible. But this strategy is not always appropriate.
>
> Let's take the problem you have just solved:
>
> All reptiles are cold-blooded.
>
> Snakes are cold-blooded.
>
> Snakes are reptiles.
>
> One way to conceptualise the logical structure of this problem is to define it like this:
>
> All A's (all reptiles) are B's (cold-blooded).
>
> The C's (Snakes) are B's (cold-blooded).
>
> So all C's (Snakes) are A's (Reptiles).
>
> In this example, it is obvious that we all know that snakes are reptiles. However, the logical structure of the problem tells us that the conclusion does not follow logically.
>
> Here, the essential point is that the premise that "All A's are B's" implies that everything that is said to be "A" is necessarily "B". But the reverse does not work. Just because something is "B" does not mean that it is also "A".
>
> If in some cases our beliefs lead us to believe that the conclusion follows logically, this is not always the case!
>
> For example, consider the following problem:
>
> All reptiles are cold-blooded.
>
> Fishes are cold-blooded.
>
> Fishes are reptiles.
>
> One way to conceptualise the logical structure of this problem is to define it like this:

All A's (all reptiles) are B's (cold-blooded).

The C's (Fishes) are B's (cold-blooded).

So all C's (Fishes) are A's (Reptiles).

In this last example, our beliefs about "fishes are reptiles" and the logical structure of the problem agree so that it is clear that the conclusion does not follow logically.

Thus, our beliefs sometimes lead us to give a correct answer, sometimes a wrong answer to a reasoning problem. This is why it is necessary to take into account the logical structure of each problem, and not only your belief.

After this explanation, participants responded to a second believable-invalid item followed by its explanation. Afterwards, participants were asked to perform the post-explanation block which consisted of the random presentation of two conflict and two no-conflict items. Only then reasoners started the second mini training block with unbelievable-valid items. Here is an example of the explanation we used for the unbelievable-valid items (e.g. "All vehicles need fuel. Bicycles are vehicles. Bicycles need fuel"):

The correct answer to the previous problem is that the conclusion follows logically.

Many people think that the conclusion does not follow logically, but this answer is wrong. Here is the explanation:

All the problems you have just solved are made up of a logical structure and knowledge to which the conclusion refers. One way to conceptualize the logical structure of the problem you have just answered is to consider that:

All A's are B's.

All C's are A's.

All C's are B's.

This logical structure always leads us to consider that the conclusion follows logically.

Syllogisms also call on your knowledge of the world. Most people only use their knowledge to decide whether the conclusion follows logically. Thus, they may respond that the conclusion follows logically if the conclusion is credible. But this strategy is not always appropriate.

Let's take the problem you have just solved:

All vehicles need fuel.

Bicycles are vehicles.

Bicycles need fuel.

One way to conceptualise the logical structure of this problem is to define it like this:

All A's (all vehicles) are B's (need fuel).

The C's (bicycles) are A's (vehicles).

So all the C's (bicycles) are B's (need fuel).

In this example, it is obvious that we all know that bicycles do not need fuel. However, the logical structure of the problem tells us that the conclusion follows logically.

The main thing is that for these problems, one should always act as if the first two sentences were true (even if we know in real life that this is not necessarily the case). If it is said that "All A's are B's", then everything that is said to be "A" must also be "B" (otherwise all A's would not be "B" and the first two sentences would not be true).

If in some cases our beliefs lead us to believe that the conclusion does not follow logically, this is not always the case!

For example, consider the following problem:

All vehicles need fuel.

Motorbikes are vehicles.

Motorbikes need fuel.

One way to conceptualise the logical structure of this problem is to define it like this:

All A's (all vehicles) are B's (need fuel).

C (motorbikes) are A (fuel).

So all the C's (motorbikes) are B's (need fuel).

In this last example, our beliefs about "Motorbikes need fuel" and the logical structure of the problem agree so that it is clear that the conclusion follows logically.

Thus, our beliefs sometimes lead us to give a correct answer, sometimes a wrong answer to a reasoning problem. This is why it is necessary to take into account the logical structure of each problem, and not only your belief.

Trial exclusion

We discarded all pre- and post-interventions trials in which participants failed to provide their initial answer before the deadline (7.2% of all these trials) or failed to pick the correct matrix in the load task (20.3% of the remaining trials), and we analysed the remaining 74.0% of all pre- and post-

intervention trials. We applied the same exclusion criterion for the pre- and post-explanations trials during the intervention in which people failed to provide their initial answer before the deadline on 8.9% of pre- and post-explanations trials or failed to pick the correct matrix in the load task on 8.6% of the remaining trials. Thus, we analysed the remaining 83.2% of all pre- and post-explanations trials. On average, each participant contributed 26.2 (SEM = 0.3) trials out of 32.

Results and discussion

Syllogism response accuracy. For each participant, we calculated the average proportion of correct initial and final responses, for conflict and no-conflict problems, in each of the two blocks (pre- and post-intervention). We analysed the data using mixed-design ANOVAs on initial and final accuracies with Block (pre- VS post-intervention) as a within-subjects factor and Group (training VS control) as a between-subjects factor.

Figure 2 shows the results. As the figure shows, on the critical conflict problems there is a tendency towards an increased post-intervention performance in the training group, both for the initial (+17.8 points) and final (+14.2 points) trials. Again, this effect was less pronounced than on previous intervention studies with numerical reasoning problems (see Study 1). The ANOVAs indicated that the central Block x Group interaction reached significance for the initial, $F(1,98) = 4.32$, $p = .04$, $\eta^2g = .01$ although not for final conflict trials, $F(1,99) = 1.55$, $p = .22$, $\eta^2g = .00$.

Nevertheless, despite the small conflict trial trend, as in Study 1, the training still seems to have led astray some participants on no-conflict control problems (see Figure 2). While participants from the control group showed stable no-conflict performance before and after the intervention, participants from the training group showed a decrease in performance after the intervention for final responses and to a lesser extent for initial responses. The ANOVAs on no-conflict accuracy showed a significant Block x Group interaction for the final trials, $F(1,100) = 8.15$, $p = .01$, $\eta^2g = .00$. The Block x Group interaction for the initial trials did not reach significance, $F(1,96) = 2.08$, $p = .15$, $\eta^2g = .01$.

Finally, we investigated the type of logical structure effect on the no-conflict performance drop of Study 2. Similar to Study 1, an identical response pattern was detected for all participants, both for the invalid and valid syllogisms (See Supplementary Material, Section C) meaning that the drop in no-conflict accuracy after the intervention was not explained by a decrease in just one given logical structure. Again, the drop in no-conflict performance is better explained by the confusion between each type of logical structure rather than the effect of one specific valid or invalid structure.

Individual level classification. We classified reasoners according to whether participants had improved or not after the intervention, using the same criterion as in Study 1.

Whereas 22% of the control participants showed improvement after the intervention (without having received any explanation), this percentage was higher in the trained group (31% of the participants who had received an explanation showed improvement). Figure 3 further shows that 32% of the control participants and 33% of the trained ones remained biased throughout the study. In addition, 36% of the control participants and 24% of the trained ones were classified as spontaneous correct reasoners (they provided a majority of correct responses in the pre and post intervention block). "Other" participants showed an inconsistent response pattern and represented 10% of the control and 12% of the trained participants.

Accuracy across training respondent categories. Figure 4 shows that, for both correct (i.e. participants who gave a majority of "01" or "11" responses both in pre- and post-intervention) and improved reasoners (i.e. participants who gave a majority of incorrect responses in pre- and a majority of "01" or "11" responses in post-intervention) from the training group, conflict accuracy increased, and the no-conflict control performance remained stable after the intervention. On the contrary, for those who did not benefit from the training, namely, the biased and the "other" reasoners, no-conflict problem accuracy dropped after the intervention. Thus, unlike in Study 1, the negative impact of training on no-conflict accuracy was not observed with those who benefited from the explanation (i.e. improved reasoners) but it was driven by those participants who did not benefit from it.

Interestingly, as shown in Figure 3, among the post-intervention conflict trials, improved reasoners gave a majority of "11" responses, namely 63% of all post-intervention trials, while they gave less "01" responses, namely 18%. This suggests that the training specifically boosts sound intuiting rather than a more efficient deliberation.

Similarly to Study 1 we quantified the proportion of improved reasoners who genuinely benefited from the training (i.e. the post intervention conflict accuracy increase is higher than the no-conflict accuracy decrease). Results showed that 88% of the improved group genuinely benefitted from the training effect (either at the intuitive and/or deliberate stage).

In sum, our revised intervention managed to truly boost performance of some trained reasoners and among these, the boost effect was applied intuitively with no need for further deliberation. However, the positive training effect is small and those who do not benefit are actually hampered on no-conflict problems.

Intervention accuracy. In Study 2, the intervention consisted in the serial presentation of two mini training sessions, one for each type of logical structure (i.e. valid and invalid). These two independent training sessions were each split into three parts: A pre- and a post-explanation part (which consisted in the presentation of two conflict and two no-conflict items), separated by an explanation part. For exploratory purposes we also examined the findings in each specific sub-section. Full results can be found in Supplementary Material, Section E.

Overall, in the valid-problem condition, trained participants showed a stronger performance increase on conflict problems than untrained (control) participants, both for initial and final responses. Performance on no-conflict problems also improved, more so in the trained group than in the untrained (control) one.

In the invalid-problem condition the picture was somewhat less clear. In the training group, while the explanation led to a rise in no-conflict-problem accuracy, conflict-problem performance remained stable (with final responses) or even dropped (with initial responses). In the untrained control group, performance dropped both for conflict and no-conflict problems, and both for initial and final responses.

In sum, the exploratory intervention block analysis indicated that when separating valid- and invalid-problems, the negative side effect of the intervention on no-conflict problems was reduced. Hence, when only one type of logical structure was specifically trained and tested in isolation, participants did not misapply it. However, when valid and invalid rules were afterward mixed within one block (i.e. in the main post-intervention block), this created confusion and prevented some participants from applying the explanations correctly.

Conflict detection. Likewise in Study 1, we calculated a conflict detection index by contrasting confidence ratings for correctly solved no-conflict items to confidence ratings for non-correctly solved conflict items. There was no indication that the training boosted conflict detection (see Table 1) at least for initial responses. The ANOVAs showed no significant interaction for initial response: $F(1,55) = 0.09$, $p = .77$, $\eta^2g = .00$ and a trend for a Group x Block interaction for final response: $F(1,50) = 3.49$, $p = .07$, $\eta^2g = .04$.

Predictive conflict detection. Similarly to Study 1, we calculated individual differences through predictive conflict detection index before the intervention between improved and biased reasoners. For initial responses, we observed a small trend indicating better conflict detection for the improved (M = 7.8%, SEM = 3.4) compared to the biased respondents (M = 0.8%, SEM = 3.0), $t(31) = 1.54$, $p = .13$, $d = .55$. This was not the case for the final response predictive conflict detection (M improved = 2.5%, SEM = 2.5; M biased = 3.8%, SEM = 2.1; $t(28) = 0.41$, $p = .68$, $d = .16$).

Study 3

Study 2 showed that a short training describing the strategy to solve valid and invalid syllogistic problems independently, helped to slightly boost the proportion of correct responses for conflict problems. Among those who benefited from the training, these correct responses were typically generated intuitively. However, there was still evidence for a negative side effect of the training. The latter led to an overall increase in the proportion of incorrect responses for easily solvable no-conflict problems (in which believability converges with validity).

In Study 2, unlike in Study 1, this negative effect of training on no-conflict trials was not linked to the benefit of the training intervention. Accordingly, whereas in Study 1 this deleterious effect was limited to improved reasoners (those who benefited from the training intervention), in Study 2 it was observed in biased reasoners (those who did not benefit from the training). The intervention used in Study 2 was thus more successful at truly helping some reasoners to improve their performance on conflict trials without creating a performance trade-off with no-conflict trials.

In Study 3, we explored whether the results of Study 2 were sustained over time. Two months after completion, all the participants from the training group of Study 2 were invited to take part in a re-test. Study 3 used the same procedure as Study 2, except that all syllogisms had a different surface content. After the pre-intervention block, participants again went through the training intervention, and they then completed a post-intervention block. This allowed us to examine whether an additional training session could help to boost participants' performance.

Method

Participants

Thirty-three participants took part in Study 3 (out of the 51 participants from the training group in Study 2; 21 females, M = 42.3 years, SEM = 2.7). The sample consisted of 11 people who were classified as biased respondents in Study 2, nine who were correct respondents, eight who were improved respondents, and the remaining five people were classified as other respondents because of their inconsistent response pattern. We compensated participants for their time at the rate of £7 per hour.

Materials and procedure

The material and procedure were the same as in Study 2. All the problems featured new contents (see Supplementary Material Section A). We used the same consecutive, isolated training for valid and invalid logical structures as in Study 2. Participants took these in the same order as they did in Study 2.

Trial exclusion

Participants failed to provide their first answer before the deadline on 5.1% of all pre- and post-intervention and on 7.6% of all pre- and post-explanation trials. They also failed to pick the correct matrix on the load task on 16.2% of the remaining pre- post-intervention trials and on 6.6% of the remaining pre- post-explanation trials. We discarded these trials and analysed the remaining trials (79.5% of all pre- and post-intervention trials and 86.4% of all pre- and post-explanation trials). On average, each participant contributed 27.4 (SEM = 0.4) trials out of 32 to the analysis.

Results and discussion

Sustainability of the training effect

To test whether the training effect sustained over time, we compared performance of the post-intervention of Study 2 (i.e. after the first training) to that of the pre-intervention of Study 3 (i.e. two months later). We also tested whether performance in the pre-intervention of Study 3 was higher than that in the pre-intervention of Study 2.

Syllogism response accuracy. For each participant, we contrasted the average proportion of correct initial and final conflict responses, across Study 2 pre-intervention, Study 2 post-intervention, and Study 3 pre-intervention blocks.

First, we focus on final-response accuracies. Figure 5 shows that participants tended to give almost as many correct responses two months after training (in the pre-intervention block of Study 3; M = 55.8%, SEM = 7.4) as directly after training (in the post-intervention block of Study 2; M = 57.8%, SEM = 7.2), $t(32) = 0.25$, $p = .80$, d = .04. Also, participants gave more correct responses two months after training (M = 55.8%, SEM = 7.4) than just before their first training (in the pre-intervention block of Study 2; M = 44.2%, SEM = 6.6), $t(32) = 1.87$, $p = .07$, d = .33. Overall, these results indicate that, for final responses, the training effect sustained over time, for at least two months after the first training.

The same trend was observed for initial responses. Performance observed two months after training (Study 3 pre-intervention: M = 48.0%, SEM = 6.5) was comparable to that observed just after training (Study 2 post-intervention: M = 54.7%, SEM = 7.1), $t(31) = 0.74$, $p = .47$, d = .13, and it was better than that observed before the first training (Study 2 pre-intervention: M = 33.3%, SEM = 5.5), $t(32) = 2.66$, $p = .01$, d = .46.

Given that not all participants of Study 2 accepted to take part in Study 3 (33/51, that is 65%), we checked for a possible attrition confound (i.e. whether those who did better in Study 2 were more likely to sign-up for Study 3). We compared the Study 2 pre-intervention conflict problem accuracy in the subgroup of participants who took part in the re-test (Initial response: M = 33.3%, SEM = 5.5; Final response: M = 44.2%, SEM = 6.6) to that for the participants who were invited to the re-test but declined to take part (Initial response: M = 36.6%, SEM = 8.7; Final response: M = 42.1%, SEM = 8.9). Given that both groups showed similar accuracy rates (Initial response: $t(49) = 0.33$, $p = .74$, d

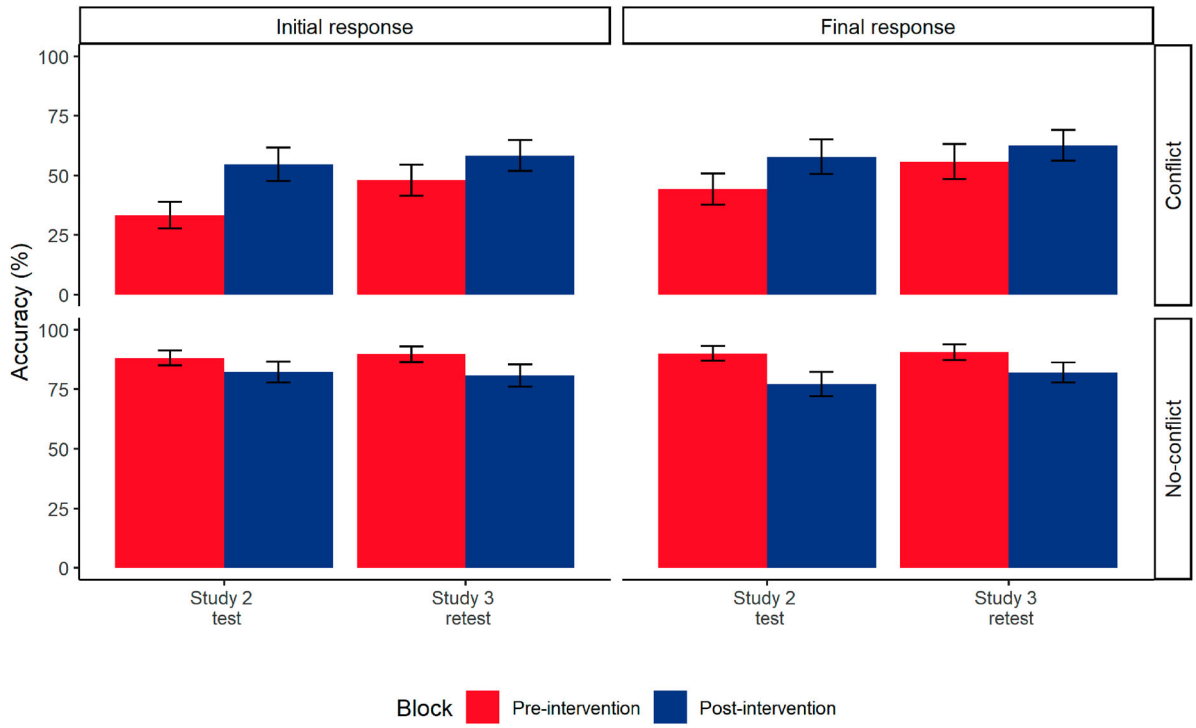

Figure 5. Average initial and final accuracies on conflict and no-conflict problems for the participants who took part to the re-test, in Study 2 (test) and Study 3 (retest). Error bars represent standard errors of mean.

= -.09; Final response: $t(49) = 0.19$, $p = .85$, d = .05), it is unlikely that the results of Study 3 are artificially boosted because of an attrition confound.

In conclusion, the training intervention effect on the conflict problems observed in Study 2 was robust and sustained over time, for at least two months, for both initial "intuitive" responses and final "deliberate" responses. This result was also supported by a direction of change analysis (see Supplementary Material Section F).

No-conflict problem accuracies were also analysed to test whether the negative training effect observed in Study 2 sustained over time. Figure 5 suggests that participants gave more correct no-conflict responses two months after the training than just after it. This finding reached significance with final (deliberate) responses, $t(31) = 2.38$, $p = .02$, d = .42, but not with initial (intuitive) accuracies, $t(29) = 1.39$, $p = .17$, d = .25. These results suggest that the negative effect of training on no-conflict response accuracies faded two months after the first training.

We also tested for a potential attrition cofound for no conflict problems. Specifically, we tested whether the participants who did better with no-conflict problems in Study 2 were more likely to sign-up for Study 3. Both groups showed similar levels of accuracy (participants who, respectively, took part and did not take part to the retest; initial response: M = 88.1%, SEM = 3.1, and M = 80.6%, SEM = 5.2, $t(49) = 1.32$, $p = .19$, d = .39; final response: M = 90.2%, SEM = 3.0, and M = 83.8%, SEM = 5.3, $t(49) = 1.13$, $p = .27$, d = .33).

Additional data. Like in Study 2, we also collected confidence ratings. We had no a priori hypotheses about these data, but the interested reader can find an overview of the results in Supplementary Material Section G.

Second training effect

In Study 3, we also tested whether a second training further improved performance. We compared performance across the pre- and post-intervention blocks of Study 3, and across the post-intervention blocks of Studies 2 and 3.

Syllogism response accuracy. First, we focus on final-response accuracies. Figure 5 shows that participants gave fairly similar levels of correct responses before (M = 55.8%, SEM = 7.4) and after the intervention of Study 3 (M = 62.6%, SEM = 6.4), $t(32) = 1.10$, $p = .28$, d = .19. The difference between Study 3 post-intervention (M = 62.6%, SEM = 6.4) and Study 2 post-intervention

performance (M = 57.8%, SEM = 7.2) did not reach significance, $t(32) = 0.74$, $p = .46$, d = .13.

With respect to initial-response accuracies, participants' performance appeared to be higher after the intervention of Study 3 (M = 58.3%, SEM = 6.5) than just before it (M = 48.0%, SEM = 6.5), but the comparison failed to reach significance: $t(32) = 1.68$, $p = .10$, d = .29. In addition, the performance observed after the intervention of Study 3 was similar to that observed after the intervention of Study 2 (M = 54.7%, SEM = 7.1), $t(31) = 0.78$, $p = .44$, d = .14. In other words, the second syllogistic training (in Study 3) did not boost performance beyond the level reached after the first training (in Study 2).

Note that the accuracy results presented here are also supported by a direction of change analysis (see Supplementary Material Section F).

Discussion

Previous research demonstrated that reasoners could be debiased with a one-shot intervention aiming at explaining the correct solution to numerical problems such as bat-and-ball, base-rate or conjunction fallacy items (Boissin et al., 2021, 2022; Bourgeois-Gironde & Van Der Henst, 2009; Claidière et al., 2017; Hoover & Healy, 2017; Morewedge et al., 2015; Purcell et al., 2020; Trouche et al., 2014). The current study tested the generalisation of these debiasing intervention findings on non-numerical syllogistic belief bias problems.

Unfortunately, the syllogistic training effect was less clear than with numerical problems. In Study 1 we failed to observe a proper debiasing effect: Despite a performance improvement on conflict problems, performance on easily solvable control no-conflict problems was impaired—especially for those participants who showed a conflict problem improvement.

Our revised training design in Study 2 was more successful at truly helping some reasoners to improve their performance on conflict trials without creating a performance trade-off on the no-conflict trials. Similar to previous studies, the training also succeed in boosting performance as early as the intuitive response stage and persisted up to two months after training (e.g. Boissin et al., 2021, 2022). However, in contrast to these previous studies which showed large effects, the current improvements were small (±10% accuracy increase) and only helped a small minority of individuals. Moreover, the negative training impact on no-conflict problems was still observed for those individuals whose conflict performance did not improve.

Overall, the findings challenge the application of the current approach as a successful debiasing method for syllogistic reasoning. This conclusion is strengthened by the fact that our study focused on fairly simple syllogisms (i.e. equivalent to the *Modus Ponens* and the *Affirmation of the Consequent* forms). These simple syllogisms did not even invoke negations (as in *Modus Tollens*, for example), nor multiple quantifiers (e.g. all, some, some), nor a logical relationship between more than three terms. Therefore, it seems safe to conclude that a short, one-shot intervention, as successfully used to debias people on numerical tasks in previous studies, can presently not be used to address belief bias in syllogistic reasoning.

However, this does not imply that we cannot debias reasoners about belief bias per se. In line with previous studies, we attempted to enhance individuals' reasoning performance with a single, short intervention that provided an easily accessible, very succinct reminder of the logical structure. In contrast to the debiasing interventions for numerical tasks, the training did not have the expected outcome. Of course, it is possible that one may obtain more success with a more extensive, repeated training that includes a more thorough schooling about the underlying logic. Likewise, the present study used one specific type of "easy fix" training that focused on giving brief explanations. In the heuristics and bias field, other types of "easy fix" interventions have been explored. Mata (2020), for example, drew reasoners' attention to the resolution-critical part of the Cognitive Reflection Test (CRT, Frederick, 2005), by underlining or highlighting the key premises. This approach proved successful and allowed reasoners to make fewer errors on subsequent problem. Other successful "easy-fix" approaches are the "consider the opposite" instruction (Adame, 2016; Hirt & Markman, 1995), reasoning about other people's reasoning (Mata et al., 2013; Reis et al., 2023) or giving additional feedback combined with explanations (Hogarth, 2001). Although these approaches have not been tested with syllogisms, it cannot be ruled out that other "easy fix" approaches might be effective in remediating people's biased belief-based thinking. Similarly, as one reviewer pointed out, it might also be possible to clarify the current explanations further by providing explanations both for conflict and no-conflict items.

We suspect that the current "easy-fix" single-shot training failure may also be linked to the nature of syllogistic belief bias problems. Previous successful interventions with numerical problems taught people about one single solution rule, such as equation solving for bat-and-ball (Boissin et al., 2021) or the ratio principle for base-rate problems (Boissin et al., 2022). Applying this rule always leads to the correct problem solution. Our syllogistic intervention, however, necessarily had to explain two different logical structures, namely a logically valid (i.e. Modus Ponens) and invalid (i.e. Affirmation of the Consequent) form on which believability can have conflicting effects (e.g. accept unbelievable valid inferences but reject unbelievable invalid inferences). It is possible that the explanation and practice of two different logical structures/rules at the same time during a short intervention creates interference and requires a different type of training.

As we noted in the introduction, the potential of simple debias interventions that can help people to reason (and even intuit) correctly is enormous. Previous publications have pointed to the successes of this approach for various reasoning problems and highlighted its prospects (e.g. Boissin et al., 2021, 2022). However, we also believe it's important to highlight its limitations and failures. The present study suggests that the single-shot "easy fix" intervention approach is currently not successful for remediating belief bias and will need further optimisation.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by the Idex Université Paris Cité ANR-18-IDEX-0001, France.

Data availability statement

Raw data can be downloaded from our OSF page (https://osf.io/rzm92/).

ORCID

Esther Boissin http://orcid.org/0000-0001-6485-7466
Serge Caparos http://orcid.org/0000-0001-6922-4449
Wim De Neys http://orcid.org/0000-0003-0917-8852

References

Adame, B. J. (2016). Training in the mitigation of anchoring bias: A test of the consider the-opposite strategy. *Learning and Motivation*, *53*, 36–48. https://doi.org/10.1016/j.lmot.2015.11.002

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

Boissin, E., Caparos, S., Raoelison, M., & De Neys, W. (2021). From bias to sound intuiting : Boosting correct intuitive reasoning. *Cognition*, *211*, Article 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

Bourgeois-Gironde, S., & Van Der Henst, J.-B. (2009). How to open the door to system 2: Debiasing the bat-and-ball problem. In S. Watanabe, A. P. Bloisdell, L. Huber, & A. Young (Eds.), *Rational animals, irrational humans* (pp. 235–252). Keio University Press.

Brisson, J., Schaeken, W., Markovits, H., & De Neys, W. (2018). Conflict detection and logical complexity. *Psychologica Belgica*, *58*(1), 318. https://doi.org/10.5334/pb.448

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., 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., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity : Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*(2), 269–273. https://doi.org/10.3758/s13423-013-0384-5

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think : When our brains detect that we are biased. *Psychological Science*, *19*(5), 483–489. https://doi.org/10.1111/j.1467-9280.2008.02113.x

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy : Perceived stereotypicality of black defendants predicts capital-

sentencing outcomes. *Psychological Science*, *17*(5), 383–386. https://doi.org/10.1111/j.1467-9280.2006.01716.x

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

Evans, J. S. B., Newstead, S. E., Allen, J. L., & Pollard, P. (1994). Debiasing by instruction : The case of belief bias. *European Journal of Cognitive Psychology*, *6*(3), 263–285. https://doi.org/10.1080/09541449408520148

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & 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

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology*, *71*(5), 1188–1208. https://doi.org/10.1080/17470218.2017.1313283

Frey, D. P., Bago, B., & De Neys, W. (2017). Commentary : Seeing the conflict: An attentional account of reasoning errors. *Frontiers in Psychology*, *8*, 1284. https://doi.org/10.3389/fpsyg.2017.01284

Gao, Q., & Liu, X. (2021). Stand against anti-Asian racial discrimination during COVID-19 : A call for action. *International Social Work*, *64*(2), 261–264. https://doi.org/10.1177/0020872820970610

Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology*, *69*(6), 1069. https://doi.org/10.1037/0022-3514.69.6.1069

Hogarth, R. M. (2001). *Educating intuition*. University of Chicago Press.

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 & Review*, *24*(6), 1922–1928. https://doi.org/10.3758/s13423-017-1241-8

Hoover, J. D., & Healy, A. F. (2019). The Bat-and-ball problem: Stronger evidence in support of a conscious error process. *Decision*, *6*(4), 369. https://doi.org/10.1037/dec0000107

Isler, O., Yılmaz, O., & Doğruyol, B. (2020). Activating reflective thinking with decision justification and debiasing training. *Judgment and Decision Making*, *15*(6), 926–938. https://doi.org/10.1017/S1930297500008147

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Strauss, Giroux.

Kahneman, D., Slovic, S. P., Slovic, P., Tversky, A., & Press, C. U. (1982). *Judgment under uncertainty : Heuristics and biases*. Cambridge University Press.

Koller, J. E., Villinger, K., Lages, N. C., Brünecke, I., Debbeler, J. M., Engel, K. D., Grieble, S., Homann, P. C., Kaufmann, R., Koppe, K. M., Oppenheimer, H., Radtke, V. C., Rogula, S., Stähler, J., Renner, B., & Schupp, H. T. (2021). Stigmatization of Chinese and Asian-looking people during the COVID-19 pandemic in Germany. *BMC Public Health*, *21*(1), 1–7. https://doi.org/10.1186/s12889-021-11270-1

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

Mata, A. (2020). An easy fix for reasoning errors : Attention capturers improve reasoning performance. *Quarterly Journal of Experimental Psychology*, *73*(10), 1695–1702. https://doi.org/10.1177/1747021820931499

Mata, A., Fiedler, K., Ferreira, M. B., & Almeida, T. (2013). Reasoning about others' reasoning. *Journal of Experimental Social Psychology*, *49*(3), 486–491. https://doi.org/10.1016/j.jesp.2013.01.010

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., 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

Newstead, S. E., Pollard, P., Evans, J. S. B., & Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition*, *45*(3), 257–284. https://doi.org/10.1016/0010-0277(92)90019-E

Nisbett, R. E. (1993). *Rules for reasoning*. Psychology Press.

Oakhill, J., Johnson-Laird, P. N., & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition*, *31*(2), 117–140. https://doi.org/10.1016/0010-0277(89)90020-6

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001

Purcell, Z. A., Wastell, C. A., & Sweller, N. (2020). Domain-specific experience and dual-process thinking. *Thinking & Reasoning*, *27*(2), 239–267. https://doi.org/10.1080/13546783.2020.1793813

Raoelison, M., & De Neys, W. (2019). Do we debias 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., Keime, M., & De Neys, W. (2021). Think slow, then fast : Does repeated deliberation boost correct intuitive responding? *Memory & Cognition*, *49*(5), 873–883. https://doi.org/10.3758/s13421-021-01140-x

Raoelison, M., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, Article 104381. https://doi.org/10.1016/j.cognition.2020.104381

Reis, J., Ferreira, M. B., Mata, A., Seruti, A., & Garcia-Marques, L. (2023). Anchoring in a Social Context: How the Possibility of Being Misinformed by Others Impacts One's Judgment. *Social Cognition*, *41*(1), 67–87. https://doi.org/10.1521/soco.2023.41.1.67

Stanovich, K. E. (2011). *Rationality and the reflective mind*. Oxford University Press.

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

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

Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2004). Everyday conditional reasoning with working memory preload. *Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society*, *26*(26), 1399–1404. https://escholarship.org/content/qt7kk1x3qx/qt7kk1x3qx.pdf

Supplementary Material

A. Items used in Study 1, Study 2 and Study 3

Syllogisms used in Study 1 and Study 2

Logical structure	Conflict	Item
Modus Ponens (valid item)	Conflict item (unbelievable)	All mammals can walk. Whales are mammals. Whales can walk.
Modus Ponens (valid item)	Conflict item (unbelievable)	All metals are solid. Mercury is a metal. Mercury is solid.
Modus Ponens (valid item)	Conflict item (unbelievable)	All vehicles have wheels. Boats are vehicles. Boats have wheels.
Modus Ponens (valid item)	Conflict item (unbelievable)	A tree will become tall. Bonsai are trees. Bonsai will become tall.
Modus Ponens (valid item)	Conflict item (unbelievable)	All birds can fly. Penguins are birds. Penguins can fly.
Modus Ponens (valid item)	Conflict item (unbelievable)	All things made of metal shine. Old pennies are made of metal. Old pennies shine.
Modus Ponens (valid item)	Conflict item (unbelievable)	All humans have two legs. Leg amputees are humans. Leg amputees have two legs.
Modus Ponens (valid item)	Conflict item (unbelievable)	All things that need oxygen have lungs. Fire needs oxygen. Fire has lungs
Affirmation of the consequent (invalid item)	Conflict item (believable)	All trees have roots. Oaks have roots. Oaks are trees.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All fruits can be eaten. Strawberries can be eaten. Strawberries are fruits.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All flowers need water. Roses need water. Roses are flowers.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All African countries are warm. Congo is warm. Congo is an African country.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All things made of wood can be used as fuel. Trees can be used as fuel. Trees are made of wood.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All sports require equipment. Hockey requires equiment. Hockey is a sport.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All dofs have snouts. Labradors have snouts. Labradors are dogs.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All things that are smoked are bad for your health. Cigarettes are bad for your health. Cigarettes are smoked.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All mammals can walk. Birds can walk. Birds are mammals.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All metals are solid. Ceramic is solid. Ceramic is a metal.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All vehicles have wheels. Trolley suitcases have wheels. Trolley suitcases are vehicles.

Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All trees will become tall. Skyscrapers under construction will become tall. Skyscrapers under construction are trees.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All birds can fly. Planes can fly. Plane are birds.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All things made of metal shine. Diamonds shine. Diamons are made of metal.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All humans have two legs. Monkey have two legs. Monkeys are humans.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All things that need oxygen have lungs. Dead people have lungs. Dead people need oxygen.
Modus Ponens (valid item)	No-conflict item (believable)	All trees have roots. Oaks are trees. Oaks have roots.
Modus Ponens (valid item)	No-conflict item (believable)	All fruits can be eaten. Strawberries are fruits. Strawberries can be eaten.
Modus Ponens (valid item)	No-conflict item (believable)	All flowers need water. Roses are flowers. Roses need water.
Modus Ponens (valid item)	No-conflict item (believable)	All African countries are warm. Congo is an African country. Congo is warm.
Modus Ponens (valid item)	No-conflict item (believable)	All things made of wood can be used as fuel. Trees are made of wood. Trees can be used as fuel.
Modus Ponens (valid item)	No-conflict item (believable)	All sports require equipement. Hockey is a sport. Hockey requires equipment.
Modus Ponens (valid item)	No-conflict item (believable)	All dogs have snouts. Labradors are dogs. Labradors have snouts.
Modus Ponens (valid item)	No-conflict item (believable)	All things that are smoked are bad for your health. Cigarettes are smoked. Cigarettes are bad for your health.
Modus Ponens (valid item)	Neutral item	Every YYY are BBB. Every AAA are YYY. Every AAA are BBB.
Modus Ponens (valid item)	Neutral item	Every KKK are DDD. Every MMM are KKK. Every MMM are DDD.
Affirmation of the consequent (invalid item)	Neutral item	Every CCC are ZZZ. Every XXX are ZZZ. Every XXX are CCC.
Affirmation of the consequent (invalid item)	Neutral item	Every RRR are EEE. Every NNN are EEE. Every RRR are NNN.
Modus Tollens (valid item)	Transfer item (believable)	All things with four legs are dangerous. Poodles are not dangerous. Poodles do not have four legs.
Modus Tollens (valid item)	Transfer item (believable)	All animals love water. Cats do not like water. Cats are not animals.
Conjunction fallacy	Transfer item	James is 26. He lives in Manhattan. He likes to wear designer clothes and acts somewhat stuck-up. On Sunday he plays golf with his father. - James volunteers in the day care center in his free time - James volunteers in the day care center in his free time and works as a stock broker
Conjunction fallacy	Transfer item	Jake is 20. He grew up in a poor family in a neglected neighbourhood. He is quite violent and already served a short sentence in prison.

		- Jake plays the violin - Jake plays the violin and is jobless

Syllogisms used in Study 2 and Study 3

Logical structure	Conflict	Items
Affirmation of the consequent (invalid item)	Conflict item (believable)	All flowers need water. Daisies need water. Daisies are flowers.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All the planets revolve around the sun. The Earth revolves around the sun. The Earth is a planet.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All animals like water. Labradors like water. Labradors are animals.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All dairy products are edible. Cheeses are edible. Cheeses are dairy products.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All birds can fly. Storks can fly. Storks are birds.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All the monuments are big. The pyramids are big. Pyramids are monuments.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All living things need oxygen. Humans need oxygen. Humans are living beings.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All humans have eyes. Blind people have eyes. Blind people are humans.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All birds can fly. Planes can fly. Plane are birds.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All things made of metal shine. Diamonds shine. Diamons are made of metal.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All humans have two legs. Monkey have two legs. Monkeys are humans.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All things that need oxygen have lungs. Dead people have lungs. Dead people need oxygen.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All fruits are eaten with dessert. The cakes are eaten with dessert. Cakes are fruit.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All beers are bitter. Endives are bitter. Endives are beers.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All pirates love gold. Gold diggers love gold. Gold diggers are pirates.
Affirmation of the consequent (invalid item)	No-conflict item (unbelievable)	All cacti have thorns. Roses have thorns. Roses are cacti.
Affirmation of the consequent (invalid item)	Conflict item (believable)	All reptiles are cold-blooded. Snakes are cold-blooded. Snakes are reptiles.
Modus Ponens (valid item)	Conflict item (unbelievable)	All mammals can walk. Whales are mammals. Whales can walk.
Modus Ponens (valid item)	Conflict item (unbelievable)	All metals are solid. Mercury is a metal. Mercury is solid.

Modus Ponens (valid item)	Conflict item (unbelievable)	All vehicles have wheels. Boats are vehicles. Boats have wheels.
Modus Ponens (valid item)	Conflict item (unbelievable)	A tree wil become tall. Bonsai are trees. Bonsai will become tall.
Modus Ponens (valid item)	Conflict item (unbelievable)	All those who wear uniforms are police officers. Firefighters wear uniforms. Firefighters are police officers.
Modus Ponens (valid item)	Conflict item (unbelievable)	All humans speak. Mute people are human. Mute people speak.
Modus Ponens (valid item)	Conflict item (unbelievable)	All plants are green. Carrots are plants. Carrots are green.
Modus Ponens (valid item)	Conflict item (unbelievable)	All humans eat meat. Vegetarians are humans. Vegetarians eat meat.
Modus Ponens (valid item)	No-conflict item (believable)	All trees have roots. Oaks are trees. Oaks have roots.
Modus Ponens (valid item)	No-conflict item (believable)	All fruits can be eaten. Strawberries are fruits. Strawberries can be eaten.
Modus Ponens (valid item)	No-conflict item (believable)	All flowers need water. Roses are flowers. Roses need water.
Modus Ponens (valid item)	No-conflict item (believable)	All African countries are warm. Congo is an African country. Congo is warm.
Modus Ponens (valid item)	No-conflict item (believable)	All felines have whiskers. Cats are felines. Cats have whiskers.
Modus Ponens (valid item)	No-conflict item (believable)	All painters are artists. Van Ghogh is a painter. Van Gogh is an artist.
Modus Ponens (valid item)	No-conflict item (believable)	All birds lay eggs. Chickens are birds. Chickens lay eggs.
Modus Ponens (valid item)	Conflict item (unbelievable)	All vehicles need fuel. Bicycles are vehicles. Bicycles need fuel.
Modus Ponens (valid item)	Conflict item (unbelievable)	All sports require equipment. Running is a sport. Running requires equipment.
Modus Ponens (valid item)	No-conflict item (believable)	All fish have fins. Sharks are fish. Sharks have fins.

B. Data for the type of justification from Study 1

In the first syllogism training study, after the last conflict problem of the post-intervention, participants were asked to select a rationale for their final response. They had to choose between four possible choices. This appeared on the screen:

> *We are interested in the reasoning behind your response to the final question:*
>
> *All things that are smoked are bad for your health.*
> *Cigarettes are bad for your health.*
> *Cigarettes are smoked.*
>
> *Does the conclusion follow logically?*
>
> *Could you please justify, why do you think that your previously entered response is the correct response to the question? Please enter your answer in the text box below:*

The coding format and procedure was based on Bago and De Neys (2019). A justification was considered as correct when it explicitly mentioned the logical structure (e.g. “Just because all things that are smoked are bad for your health, doesn't mean all things that are bad for your health can be smoked, so just because cigarettes are bad for your health, it doesn't conclude that they can be smoked.”). All other responses were coded as incorrect.

Table S1.

Frequency of different types of justifications for the final syllogistic conflict problem during the post-intervention in Study 1.

Justification	**Control group**		**Training group**	
	Correct response	*Incorrect response*	*Correct response*	*Incorrect response*

	(n = 19)	*(n = 25)*	*(n=19)*	*(n=26)*
Correct	13	2	10	1
Incorrect	6	23	9	25

Note. Justification data of 10 participants is missing because their trial was excluded due to a missed deadline (see Exclusion Criteria).

C. Accuracy of each logical structure on pre- and post-intervention of Study 1 and Study 2

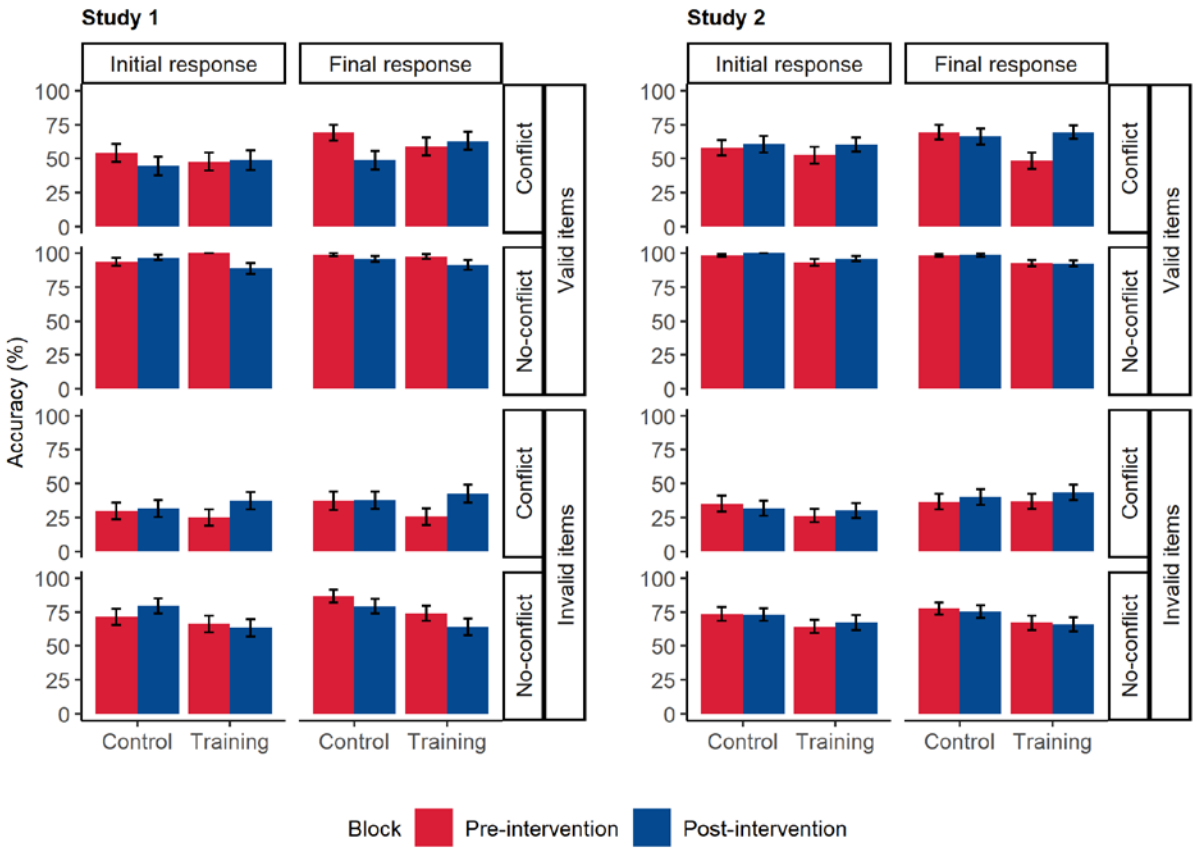

Figure S1. Average initial and final response accuracies on conflict and no-conflict valid and invalid problems in Study 1 and 2, for each group (i.e., Control VS. Training), before and after the intervention. Error bars represent standard errors of the mean (SEM).

D. Neutral and Transfer problems accuracies in Study 1

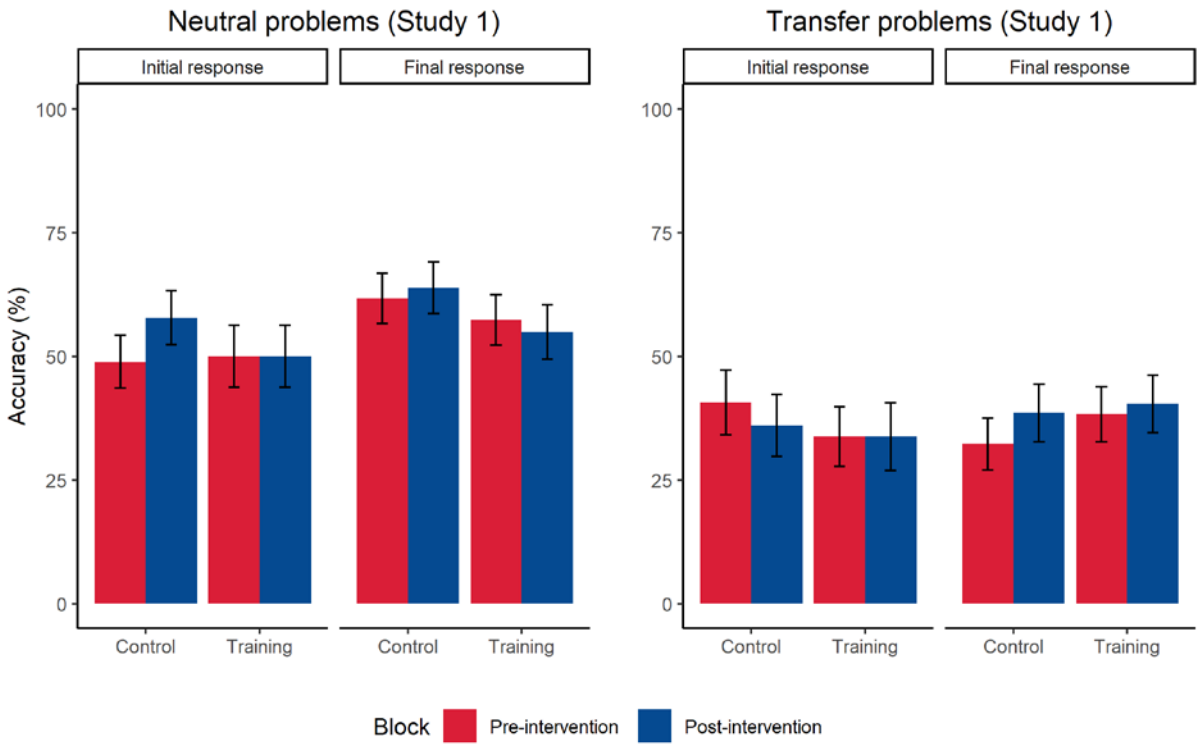

Figure S2. Average initial and final accuracy on neutral and transfer problems in Study 1 across Groups, before and after the intervention. Error bars represent standard error of the mean (SEM).

E. Mini training intervention accuracies during the pre- and post-explanation in Study 2

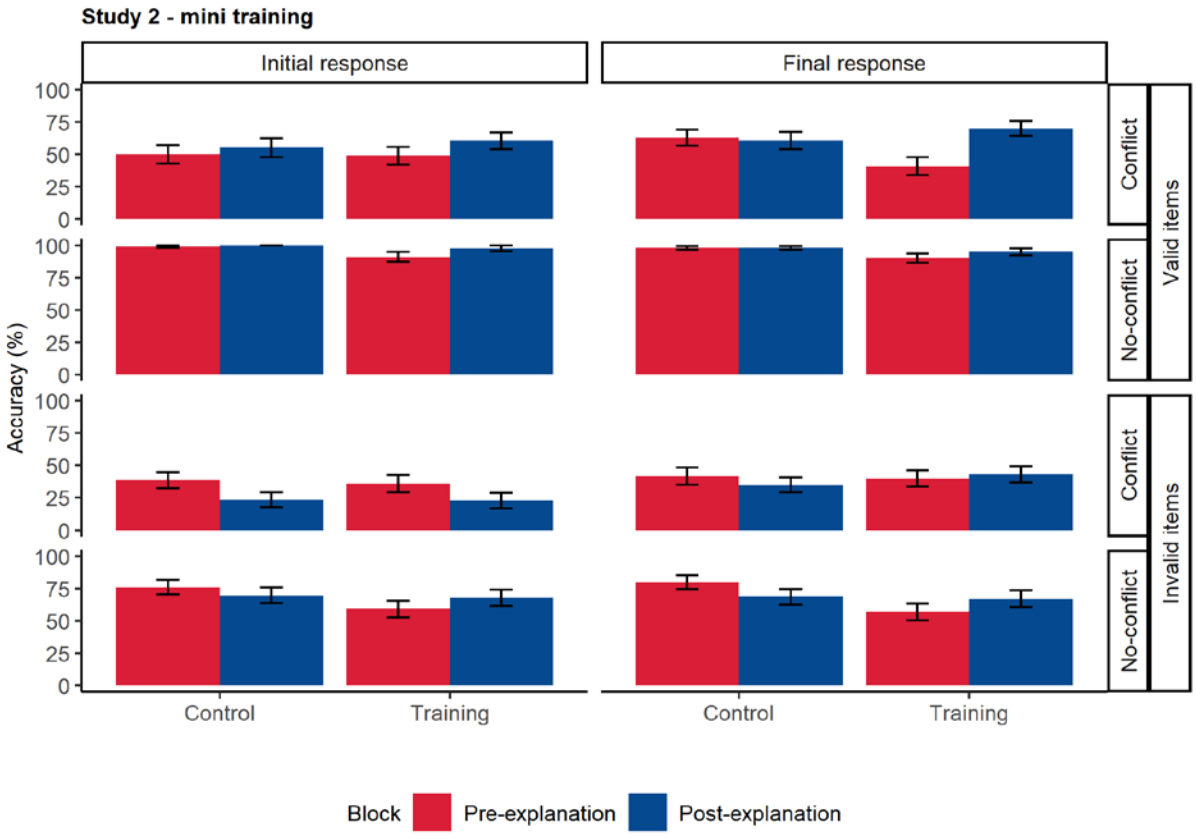

Figure S3. Average initial and final response accuracies on conflict and no-conflict for valid and invalid problems in Study 2, for each group (i.e., Control VS. Training), before and after the explanation of each mini training (valid VS invalid). Error bars represent standard errors of the mean (SEM).

F. Direction of change analyses of Study 2 (test) and Study 3 (re-test)

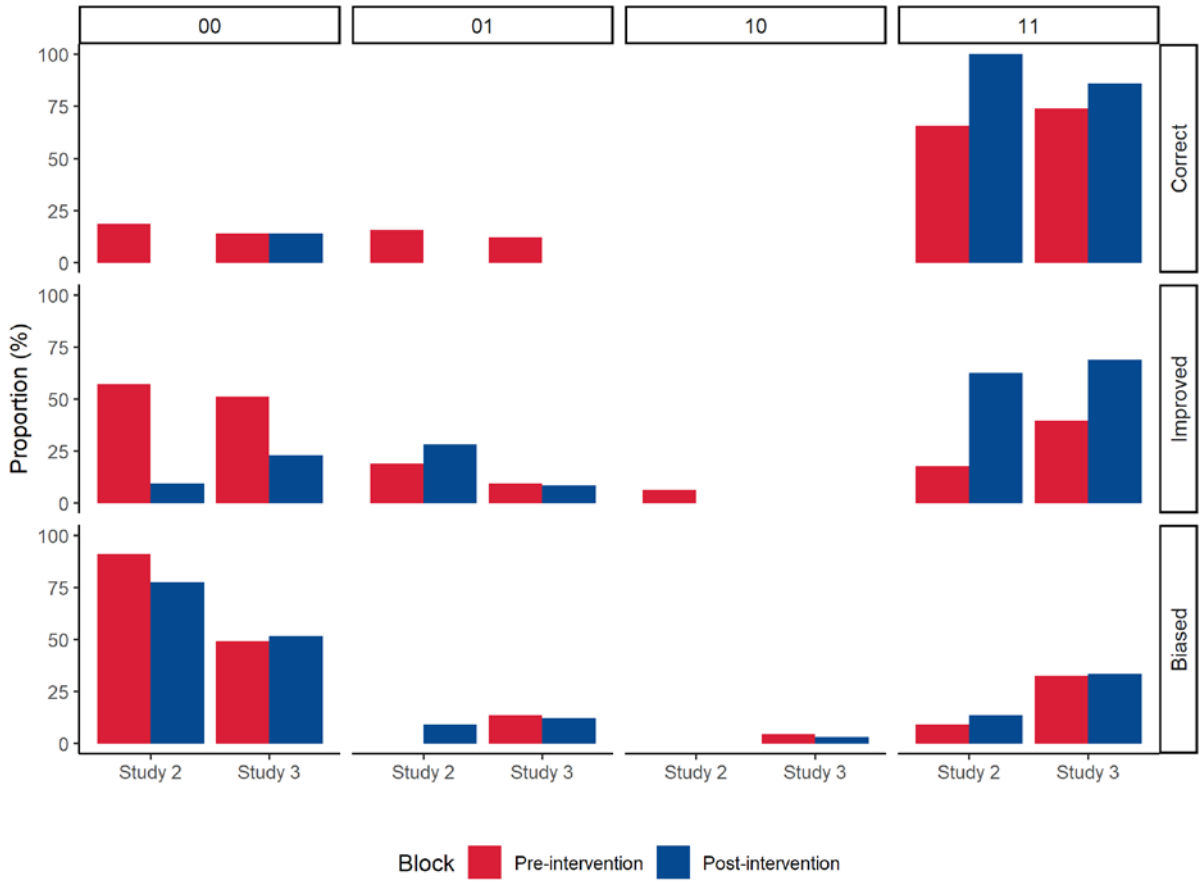

Figure S4. Proportion of each direction of change (i.e., 00 trials, 01 trials, 10 trials and 11 trials) for the conflict problems according to Block (Pre-intervention VS Post-intervention of the participants who took part in the-rest (Study 3) compared to two months before (Study 2).

G. Conflict detection index with confidence ratings for Study 3

Table S2.

Percentage of mean differences in confidence ratings (SEM) between conflict and no-conflict problems as an index of conflict detection.

Block	Initial response	Final response
Pre-intervention	-11.7 (4.9)	-0.1 (5.2)
Post-intervention	-2.9 (5.6)	4.6 (6.4)