Boosting debiasing: Impact of repeated training on reasoning

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

Boosting Debiasing: Impact of Repeated Training on Reasoning

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

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

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

Abstract

Background: Recent debiasing studies have shown that a short explanation about the correct solution to a reasoning problem can often improve performance of initially biased reasoners. Yet, with only one single training session, there is still a non-neglectable group of reasoners who remained biased.

Aims: We explored whether repeated training on a battery of three reasoning tasks (i.e., bat-and-ball, base-rate neglect, and conjunction fallacy) can further boost reasoning performance.

Sample: We recruited 120 adults, native English speakers, through Prolific Academic.

Methods: We ran two studies with a battery of three classic reasoning tasks (see above). We used a two-response paradigm in which participants first gave an initial intuitive response, under time pressure and cognitive load, and then gave a final response after deliberation. In Study 1, we ran two repeated training sessions within one week. In Study 2, we ran a third training session two months after the initial study.

Results: Study 1 showed that after the first training session, most of the participants solved the problems correctly, as early as the initial intuitive stage. This training effect was further boosted by additional training, which helped almost the full sample to benefit. Study 2 indicated that these effects were robust and persisted after two months.

Conclusions: The repetition of the training can further boost performance compared to the effect of one single training. These results are consistent with the wider literature on repeated testing and can serve as a proof-of-principle for a repeated debias training approach.

***Keywords*:** Reasoning · Dual-process theory · Heuristics · Debiasing · Repeated training

Introduction

Decades of reasoning and decision-making research have shown that human judgment is often biased. In general, people tend to over-rely on fast intuitive impressions rather than on more demanding logico-mathematical principles (e.g., Evans, 2008; Kahneman, 2011; Stanovich & West, 2000). This intuitive or so-called "heuristic" thinking can sometimes conflict with traditional logical or probabilistic considerations in a wide range of situations (Kahneman & Frederick, 2002).

The conjunction fallacy problem, initially presented by Tversky and Kahneman (1983), illustrates this phenomenon: Imagine that you are attending a party, and you are introduced to Maddy. Through the discussion, you learn that she has previously studied gastronomy and likes French food. If you had to guess Maddy's job, would she most likely be: *'A gardener'* or *'A gardener and a wine taster'*? For many of us, the first intuitive response that spontaneously springs to mind is '*A gardener and a wine taster'*, because this is the response that best fits with our idea of someone that has studied gastronomy and likes French food. However, the cued stereotypical association violates the conjunction rule, which stipulates that the probability of a conjunction, P(A&B), cannot exceed the probabilities of its constituents, P(A) and P(B) (i.e., p(A&B) ≤ p(A), p(B)). That is, there will always be more individuals that are simply gardeners than individuals that are gardeners and in addition wine tasters.

A famous explanation for this biased thinking has been given by the influential dual process model, which characterizes human reasoning as an interplay between two types of processes or "systems": A fast, intuitive one (often called "System 1") and a slower, more effortful, deliberative one (often called "System 2"; e.g., Evans & Stanovich, 2013; Kahneman, 2011). Reasoners who manage to solve the problem correctly in line with standard logic or probabilistic principles (i.e., select '*A gardener*' in the above example) would correct their initially generated intuitive response (i.e., '*A gardener and a wine taster*') after completing deliberative calculations (e.g., Kahneman, 2011; Morewedge & Kahneman, 2010). However, because reasoners tend to minimize demanding computations, they will often apply the intuitive processes by default and stop there, without considering that the correct answer could be different (Evans & Stanovich, 2013; Kahneman, 2011; Kahneman & Frederick, 2005). Consequently, most reasoners remain biased.

Not surprisingly, reasoning scholars have long been trying to remediate people's biased thinking (e.g., Habib & Cassotti, 2015; Lilienfeld et al., 2009; Milkman et al., 2009). Recent successful debiasing studies have shown that a single-shot, plain-English intervention can often help people to reason more accurately (e.g., Boissin et al., 2021, 2022; Claidière, et al., 2017; Hoover & Healy, 2017; Morewedge et al., 2015; Purcell et al., 2020; Trouche et al., 2014). Typically, this intervention consists of an explanation about the correct solution strategy and the typical biased response (see 2.1.3 Materials for a full example). Once the problem has been properly explained, many initially biased reasoners manage to produce correct responses to structurally similar problems afterwards.

Such results are obviously promising. However, the nature of the training effect is currently not clear. A key question is whether the training primarily affects people's intuitive or deliberate thinking. The common assumption is that after training, participants will be more likely to deliberate properly (i.e., to engage their "System 2") and correct the intuitively generated heuristic response (e.g.,

Evans, 2019; Lilienfeld et al., 2009; Milkman et al., 2009). This assumption fits with the general dual process idea that the deliberate "System 2" primarily serves to correct the intuitive "System 1" (e.g., Kahneman, 2011; Pennycook et al., 2015b). However, in theory, it is also possible that once reasoners grasp the solution, they will no longer generate an incorrect intuitive response. Instead, they might intuitively apply the correct solution strategy without the need for a corrective "System 2" deliberation process.

If a debiasing training actually helps people intuit correctly, this would have far-stretching implications (see Boissin et al., 2021, 2022). Although it can be laudable to help people deliberate more, in many daily life situations they will simply not have the time (or resources/motivation) to deliberate. Hence, as Boissin et al. (2021) put it, if debiasing interventions only help people to deliberately correct erroneous intuitions, their impact may be suboptimal. Ultimately, we do not only want people to learn to correct erroneous intuitions, but to avoid biased intuitions altogether (Milkman et al., 2009; Reyna et al., 2015). The potential benefits of training sound intuitions are rife in this respect.

Recent evidence lends some credence to the "trained intuitor" viewpoint (e.g., Boissin et al., 2021, 2022). These studies used a two-response paradigm (Thompson et al., 2011) to determine whether the explanation affected participants' intuitive and/or deliberate reasoning. In this paradigm, reasoners are asked to give two consecutive responses to a given problem. First, they have to provide their initial "intuitive" response under time-pressure and, at the same time, perform a secondary memory-task that burdens people's cognitive resources and disrupts the potential involvement of the deliberative "system" (Bago & De Neys, 2019). Immediately afterwards, they are presented with the problem again and can take all the time they need to think about it and give their final "deliberate" response. Two-response findings indicate that whereas most reasoners were biased before the training (both at the initial and final response stages), immediately after the explanation intervention most of them are able to provide correct responses. Critically, their responses were correct as early as the intuitive stage (Boissin et al., 2021, 2022). This suggests that the explanation debiasing approach allows people to intuit correctly (rather than to boost their deliberate correction).

Given that the "sound-intuiting" debiasing approach has important applied and theoretical implications, further validation for the "trained intuitor" viewpoint is needed. This is especially crucial since even though most reasoners benefited from Boissin et al.'s (2021, 2022) debiasing training, a non-neglectable group remained biased. However, as many psychological training studies, Boissin et al. (2021, 2022) only presented their participants with a single training session. It has long been argued in the educational and learning field that more frequent training sessions might boost learning and maintain acquired knowledge (e.g., see Rawson & Dunlosky, 2022). Hence, multiple training sessions

may help yield better and more enduring effects compared to a single shot training session (e.g., Benjamin & Tullis, 2010; Carpenter et al., 2022; Higham et al., 2022; Karpicke & Bauernschmidt, 2011). The present work aims to test the impact of a repeated debiasing training on participants' reasoning performance.

In Study 1, we measured the short-term impact of a second training: After a first training (*Session 1*), participants were invited to a second training session two days later (*Session 2*). In Study 2, we re-tested trained participants from Study 1 two months after the initial training (*Session 3*) to explore whether the training effect was robust and sustained over time. In each session, we tested the reasoning performance of participants on three notorious reasoning tasks: The bat-and-ball (Frederick, 2005), base-rate neglect (Kahneman & Tversky, 1973) and conjunction fallacy tasks (Tversky & Kahneman, 1983). They were combined in a one-hour training battery. This allowed us to test the generalizability of the training effect across different reasoning tasks. For each task, the training took around 20 minutes and consisted of three different blocks: A pre-intervention, an intervention, and a post-intervention. Participants were assigned to a training or control group. In the intervention block, participants from the training group solved task problems and always received a short debiasing explanation about the rationale behind the task, while participants of the control group simply solved the problems without receiving the explanation. During the pre- and post-intervention blocks, we used the two-response paradigm to determine whether the intervention affected participants' intuitive and/or deliberate reasoning.

To avoid confusion, we also present some clarification with respect to our nomenclature. In line with the debiasing literature, we use different concepts such as bias, logical fallacies, and heuristics interchangeably. One could argue that our interest concerns more specifically logical fallacies in classic reasoning tasks. Debiasing refers to an accuracy improvement on these tasks. Note that this does not imply that the use of heuristics is necessarily problematic. In many contexts, they may provide valid problem solutions. We specifically focus on classic reasoning tasks from the heuristics and biases literature that are designed such that cued heuristics conflict with elementary logical principles. Hence, consistent with previous training studies (e.g., Boissin et al., 2021, 2022; Hoover & Healy, 2017), if people can learn to discard erroneous mathematical or stereotypical heuristics after an intervention, we refer to this as a "debiasing" effect.

Study 1

Method

Preregistration and data availability

The study design and research questions were preregistered on the AsPredicted website (https://aspredicted.org) and stored on the Open Science Framework. No specific analyses were preregistered. All data and analysis scripts are also available on the Open Science Framework (http://osf.io/3aqh4).

Participants

Participants were recruited online, using the Prolific Academic website (http://www.prolific.co). Only native English speakers from Canada, Australia, New Zealand, the USA, or the UK were allowed to take part in the study. Participants were informed that there would be two test sessions two days apart when signing up. They were re-contacted two days after the first test session (*Session 1*). They were paid £5 for their participation in Session 1, and £6 for their participation in Session 2.

In total, 120 reasoners participated in Session 1 (92 female, M_{age} = 37.5 years, SD = 13), 74 participants were randomly assigned to the training group and 46 to the control group[1]. Among them, 2 participants had not completed secondary school, 52 had secondary school as their highest level of education, and 66 reported a university degree. In Session 2, 110 participants out of 120 (i.e., 91.7%) took part in the re-test (86 females, M_{age} = 38.1 years, SD = 12.7). The sample was composed of 67 participants in the training group, and 43 in the control group.

Our sample size decision was based on Boissin et al.'s (2021) original study who tested 100 participants. We factored in a possible 20% attrition rate between test sessions and consequently aimed to recruit 120 participants. All reported results and analysis concern the 120 participants that completed Session 1 and the 110 participants that completed again Session 2.

Materials

Each session was composed of three different reasoning tasks (i.e., bat-and-ball, base-rate neglect, and conjunction fallacy tasks). In each session, for each participant, the task order was randomized. Each task contained eight conflict and eight no-conflict problems (see further) and was composed of three blocks presented in the following order: A pre-intervention, a short intervention, and a post-intervention block. In total, each participant had to solve 48 problems in Session 1, and

[1] Due to a coding error, more participants were allocated to the training group.

again the same number of problems in Session 2. All these problems are presented in the Supplementary Material Section A.

Bat-and-ball problems (BB). In Sessions 1 and 2, we presented problems taken from Raoelison and De Neys (2019) and Boissin et al. (2021). They were modified versions of the original bat-and-ball problem (Frederick, 2005) which used quantities instead of prices. They were presented using a free-response format, where participants typed in their response using a computer keyboard (e.g., see Bago & De Neys, 2019). In the standard conflict version of these problems, the intuitively cued heuristic response hints an answer that conflicts with the correct logical answer. For instance, in a typical conflict version ("*A city has acquired 430 buses and trains in total. There are 400 more buses than trains. How many trains are there?*"), the cued heuristic response (i.e., "*30 trains*") conflicts with the correct logical response (i.e., "*15 trains*" as 430 in total – 400 buses / 2 = 15). To assure that possible correct or incorrect responses did not originate from guessing, we also presented no-conflict control problems. In these control problems, the conflict was removed by deleting the critical relational "more than" statement. The heuristic intuition thus cued the correct response (e.g., "*A city has acquired 610 buses and trains in total. There are 600 buses. How many trains are there in this city?"*; De Neys, Rossi & Houdé, 2013). Note that, as Boissin et al. (2021), we added three words to the control problem questions to equate the semantic length of the conflict and no-conflict versions. We presented four conflict and four no-conflict control problems in the pre- and post-intervention blocks. These control problems should be easy to solve. If participants are paying minimal attention to the task and refrain from random guessing, accuracy should be at ceiling (Bago & De Neys, 2019).

Base-rate neglect problems (BR). In Sessions 1 and 2, each participant was also presented with base-rate problems taken from Bago and De Neys (2017). Participants always received a description of the composition of a sample (e.g., "This study contains high school students and librarians"), a description that was designed to cue a stereotypical association (e.g., "This person is loud") and a base rate information (e.g., "There are 5 high school students and 995 librarians"). Participants' task was to indicate to which group the person most likely belonged. The task instructions stressed that the person was drawn randomly from the specified sample. The problem presentation format was based on Pennycook et al.'s (2014) rapid-response paradigm. The base rates and descriptive information were presented serially and the amount of text presented on screen was minimized. As in Pennycook et al. (2014), base rates varied between 995/5, 996/4, and 997/3. The following illustrates the full problem format:

This study contains high school students and librarians.

Person 'D' is loud.

There are 5 high school students and 995 librarians.

Is Person 'D' more likely to be:

- *A high school student*
- *A librarian*

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

We presented four conflict and four no-conflict problems in the pre- and post-intervention blocks. In the no-conflict control problems, the description triggered a stereotypical trait of a member of the largest group. As in the other tasks, these no-conflict problems should be easy to solve. If participants are paying minimal attention to the task and refrain from random guessing, they should show high accuracy (Bago & De Neys, 2019).

Conjunction fallacy problems (CF). In Sessions 1 and 2, we used the conjunction task format introduced by Andersson et al. (2020) as adopted by Boissin et al. (2022). All conjunction problems presented a short personality description of a character, consisting of his name (e.g., "Falon"), his age (e.g., "26"), his previous studies (e.g., "education") and his hobby/interest (e.g., "children"). Next, the participants were given four response options and were asked to indicate which one was most likely. In the critical conflict problems, one option presented a characteristic that featured an unlikely stereotypical association given the description (e.g., "a flight attendant") and one option presented a conjunction of this unlikely and a likely characteristic (e.g., "a flight attendant and a dad"). Two other filler options presented a characteristic that was very unlikely (e.g., "a duke") and a conjunction of two unlikely characteristics (e.g., "a flight attendant and a rally racing fan"). The following illustrates the full problem format:

Falon, 26, has previously studied education and likes children.

Is it most probable that the described person is:

- *A flight attendant*
- *A flight attendant and a dad*
- *A duke*
- *A flight attendant and a rally racing fan*

We presented four conflict and four no-conflict control problems in the pre- and post-intervention blocks. In the no-conflict control problems, we replaced the singular unlikely response option with the option that featured the likely stereotypical association (e.g., “A dad” in the above example). Reasoners will tend to select the statement that best fits with the stereotypical description (Tversky & Kahneman, 1983). Clearly, the fit will be higher for the likely than the unlikely characteristic with the conjunctive statement falling in between. Hence, on the no-conflict problems, stereotypical associations will no longer favour the conjunctive over the singular statement and participants are expected to show high accuracies (see De Neys et al., 2011).

The four response options were presented in random order. Note that Andersson et al. (2020) adopted the four options design to minimize the use of simple visual response strategies (e.g., "always choose the shortest answer"). As in the Andersson et al. study, selection of the filler options was overall very rare in our studies (i.e., 6.3% of options in Session 1 and 5.1% of options in Session 2). However, strictly speaking, participants who select the singular very unlikely option (e.g., “a duke” in the above example) do not violate the critical conjunction rule. As Boissin et al. (2022) mentioned, given that we are interested in learning effects, selection of the very unlikely option can be considered a correct response. Hence, we considered answers on which the conjunction fallacy is avoided (i.e., unlikely and very unlikely answers) as correct answers. Figures S2 and S3 in Supplementary Material Section C give a detailed overview of the selection frequency of each individual response option.

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

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

“The correct answer to the previous problem is that Maddy is most likely "a gardener". Many people think that the answer is "a gardener and a wine taster" but this is wrong.

Most people base their answer on the description. Sometimes the description can lead us to give a correct answer, but it can also mislead us. Indeed, if we refer to Maddy's educational background and interests, it seems more realistic to think of Maddy as "a gardener and a wine taster" rather than only “a gardener”. Simply because adding that Maddy is also "a wine taster" is more in line with our representation of someone who has studied gastronomy and likes French food, rather than Maddy only being "a gardener". If one of the proposed answers would have been “a wine taster" then this reasoning would probably be correct. However, in this problem the option "a wine taster" is presented together with another event, "a gardener".

Now the statistical probability that Maddy is "a gardener" is higher than the probability that Maddy is "a gardener AND a wine taster". This is because a single event is always more probable than the combination of this event with another one, whether you think it fits the description or not.

To illustrate this reasoning, consider the category corresponding to "a gardener". Some gardeners will also be wine tasters, others will not be wine tasters. The group of people who are “gardeners and wine tasters” is a subgroup of the group of all gardeners. Hence, there will always be more people who are simply gardener than people who are gardener and in addition

also wine taster. Simply because one is a subgroup of the other, it will always be more probable that someone is a gardener rather than a gardener and a wine taster."

Two-response format. We used the two-response paradigm (Thompson et al., 2011) for the presentation of all problems in the pre- and post-intervention blocks. In this paradigm, participants are asked to provide two consecutive responses on every trial: A "fast" response, directly followed by a second "slow" response. This method allowed us to capture both an initial "intuitive" response, and then a final "deliberate" one. To minimize the possibility that deliberation was involved in producing the initial "fast" response, participants had to provide their initial answer within a strict time limit while performing a concurrent cognitive load task (e.g., Bago & De Neys, 2017, 2019; Boissin et al., 2021, 2022). The load task was based on the dot memorization task (Miyake et al., 2001) given that it had been successfully used to burden executive resources during reasoning tasks (e.g., De Neys, 2006; Franssens & De Neys, 2009). Participants had to memorize a complex visual pattern (i.e., a 3 x 3 grid in which 4 dots were placed) that was presented briefly before each reasoning problem. After their initial "intuitive" response to the problem, participants were shown four different matrixes, and they had to choose the correct pattern (see De Neys, 2006, for more details). They received feedback as to whether they chose the correct or incorrect pattern.

For all base-rate problems, a time limit of 3 seconds was chosen for the initial response, based on previous pre-testing that indicated it amounted to the time needed to read the preambles, move the mouse, and click on a response option. Similarly, the time limit was set to 8 seconds for the free-response bat-and-ball problems and 5 seconds for the four-option conjunction fallacy problems. For all tasks, previous pretesting established that the time limits imposed a stringent time-pressure that forced participants to respond significantly faster than in a traditional unconstrained, one-response test format (Bago & De Neys, 2017, 2019; Boissin et al., 2022). The debiasing studies of Boissin et al. (2021, 2022) with these three tasks used the exact same deadlines. Note that the time limit and cognitive load were only applied during the initial response stage and not during the subsequent final stage in which participants were allowed to deliberate.

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

The coding format and procedure was based on Bago and De Neys (2019) for bat-and-ball and Boissin et al. (2022) for base-rate tasks. A justification was considered as correct when it explicitly mentioned the correct calculation for the bat-and-ball (e.g., "150 in total - 100 men = 50 women / 2, the response is 25") or the use of the base-rate (e.g., "Greater number of nurses to artists. For every 1 artist there are 332 nurses, so the odds are stacked against it being an artist."). Other justifications, whether they mentioned an incorrect calculation or unspecified statement (e.g., "I did it in my head") were coded as incorrect. For the conjunction fallacy task, we adopted a criterion involving a similar procedure as above: A justification was considered as correct when it explicitly referred to the conjunction principle (e.g., "There are always more people who are simply female than female and architects"). Likewise, all other types of justifications were considered as incorrect. Because of a coding error, the base-rate justifications were not accurately recorded in Session 2 and were removed from the analysis.

Session 1 results indicated that, for the three tasks, the majority of correct responses was correctly justified after training (training group: 112 correct justifications out of 171 correct responses, i.e., 66%; control group: 26 correct justifications out of 44 correct responses, i.e., 59%). This was also the case for bat-and-ball and conjunction fallacy tasks in Session 2 (training group: 66 correct justifications out of 109 correct answers, i.e., 60%; control group: 12 correct justifications out of 13 correct responses, i.e., 92%). The interested reader can find details in Tables S8 and S9 in Supplementary Material Section I. Note that the justification was untimed and retrospective. It was collected for exploratory purposes and does obviously not allow drawing any conclusions with respect to the intuitive or deliberate nature of participants' processing.

Procedure

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

The overall procedure of a typical trial consisted of, first, presentation of a fixation cross displayed during 2000 ms, followed by the first sentence of the problem (e.g., *"In a store one can choose between 320 tomatoes and avocados"* for the bat-and-ball task) for 2000 ms, and followed by the visual matrix for the cognitive load task for 2000 ms. Then, the full problem was presented, at which point participants had 8000 ms (bat-and-ball), 3000 ms (base-rate neglect) or 5000 ms (conjunction fallacy) to give their initial answer. Note that, in this initial "intuitive" response stage, the background of the screen turned yellow after 6000 ms (bat-and-ball), 2000 ms (base-rate neglect) or 3000 ms (conjunction fallacy), to warn participants that they only had a short amount of time left to answer. If they had not provided an answer before the time limit, they were given a reminder that it was important to provide an answer within the time limit on subsequent trials. Participants were then asked to enter their confidence in the correctness of their answer on a scale from 0% (absolutely not confident) to 100% (absolutely confident). Then, they were presented with four visual matrix options and had to choose the one that they had previously memorized. Finally, the same reasoning problem was presented again, and participants were asked to provide a final "deliberate" answer (with no time limit nor cognitive load) and, once again, to indicate their confidence level.

Two days after Session 1, participants were invited for a second similar one-hour training session, composed of the same three reasoning tasks (i.e., bat-and-ball, base-rate, and conjunction fallacy tasks). The only difference with the first training was the material used: All the problems featured modified contents (see Supplementary Material Section A). The minimal two-day delay was chosen for mere practical organisational reasons. We accepted participations upon one week after launching Session 2. However, the vast majority of participants took Session 2 on the launch day (mean days delay between Sessions = 3.3, SD = 0.9). After Session 2, at the end of Study 1, participants in the control group were also presented with the explanations about how the bat-and-ball, base-rate neglect, and conjunction fallacy problems could be solved, and all participants were asked to complete a page with demographic questions.

Trial exclusion

Following our preregistration, in Session 1 and Session 2 we discarded trials in which participants failed to provide their initial answer before the deadline (2.3% of all Session 1 trials and 4.8% of all Session 2 trials) or failed to pick the correct matrix in the load task (12.0% of the remaining trials in Session 1 and 13.3% of the remaining trials in Session 2), and we analysed the remaining 86.0% of all Session 1 trials and the remaining 82.5% of all Session 2 trials. On average, each participant contributed 40.9 (SD = 5.7) conflict trials out of 48 and 41.1 (SD = 5.8) no-conflict trials out of 48 in

Session 1, and 41.1 (SD = 6.2) conflict trials out of 48 and 42.0 (SD = 7.4) no-conflict trials out of 48 in Session 2.

Note that as part of our procedure, in Session 1, we asked participants whether they were familiar with the original bat-and-ball problem (Frederick, 2005). In total, 45 participants out of 120 (37.5%) reported having come across the problem before. Traditionally, these participants are removed from the analyses to eliminate the possibility that their prior knowledge of the correct solution affects the results (e.g., Bago & De Neys, 2019; Boissin et al., 2021). First, we ran all analyses while including these 45 participants, and second, while not including them. None of our conclusions were affected either way, and the trends remained the same. Thus, in line with our preregistration, we take these participants into account in the reported analyses in the main text (see Figure S1 in Supplementary Material Section B for overview analyses with and without these participants).

Analysis strategy

For simplicity and to maximize power, our analyses focused on the composite conflict accuracy across the three different reasoning tasks (i.e., bat-and-ball, base-rate neglect, and conjunction fallacy). To calculate the composite performance, we averaged for each participant the proportion of correct initial and final responses, separately for each task. Then we averaged across all tasks (separately for initial and final trials). A correlation table of all variables of each condition can be found in Supplementary Material Section K. Cronbach alpha of the composite measure of the post-intervention trials reached .82 for initial responses and .78 for final responses (it was also computed for each of the individual tasks and varied between .79 and .81 for initial responses and .73 and .85 for final responses). The corresponding composite Cronbach alpha of the pre-intervention trials reached .79 for initial responses and .75 for final responses (for each of the individual tasks it varied between .78 and .82 for initial responses and .73 and .85 for final responses, see Supplementary Material Section K for further details). For completeness, we calculated the composite performance also for no-conflict trials (Supplementary Material Section D).

The data were processed and analysed using the R software (R CoreTeam, 2017) and the following packages (in alphabetical order): dplyr (Wickham et al., 2020), ggplot2 (Wickham, 2016), lmerTest (Kuznetsova et al., 2020) and tidyverse (Wickham, 2022).

Throughout the article, we used mixed-effect regression models in which participants were entered as random effect intercept. The Wald test assessed the statistical significance of the fixed effect of the model. Note that we tried to design a more complete model, in which both participants and items were entered as random effect intercepts. However, it failed to converge, thus we kept the simpler model described above.

Results

We will first present the accuracy results (i.e., the average proportion of correct initial and final responses, in each block and each group) of the first training session to see whether we replicate the training effects observed in the recent debiasing literature. Next, we will focus on the impact of repeated training on the response accuracy in Session 2. Finally, we will present additional analyses on the confidence data in Sessions 1 and 2.

Session 1 accuracy

Conflict trials accuracy. For each participant and for each reasoning task, we calculated the average proportion of initial "intuitive" and final "deliberate" correct responses on all conflict items. Eyeballing Figure 1 indicates that participants were typically biased and showed low final accuracies before the intervention, in both control and training groups (respectively M = 24.0%, SD = 21.9 and M = 32.7%, SD = 25.6). However, the average proportion of correct final responses improved after the intervention. Notably, they sharply increased in the training group (+44.2 points, reaching M = 76.9%, SD = 26.6) whereas they improved slightly in the control group (+8.9 points, reaching M = 32.9%, SD = 21.2). Statistical composite analyses revealed that the Block x Group interaction significantly improved the model for the final responses, $\chi 2$ (1) = 32.12, $p < .001$.

Similarly, initial accuracies also showed that reasoners typically failed to provide a correct answer before the intervention, in both control and training groups (respectively, M = 18.7%, SD = 19.1 and M = 27.8%, SD = 23.1)**.** However, initial performance also increased after the intervention. Once again, this improvement was much better in the training group (+42.0 points, reaching M = 69.8%, SD = 31.0) than in the control group (+7.9 points, reaching M = 26.6%, SD = 22.2). Statistical composite analyses indicated that the Block x Group interaction also significantly improved the model for the initial responses, $\chi 2$ (1) = 34.72, $p < .001$.

For completeness, Figure 1 (bottom panels) also shows the data for each individual reasoning task. As the figure indicates, by and large, similar initial and final response trends were observed on each of the individual reasoning tasks. If anything, the training effect tended to be somewhat less pronounced for the base-rate task, but participants' pre-intervention performance on this task was also already higher than for the others.

Overall, these results are consistent with the recent literature (e.g., see Boissin et al., 2021, 2022; Purcell et al., 2022) and confirm that a single training can significantly increase response accuracy, as early as the initial response stage.

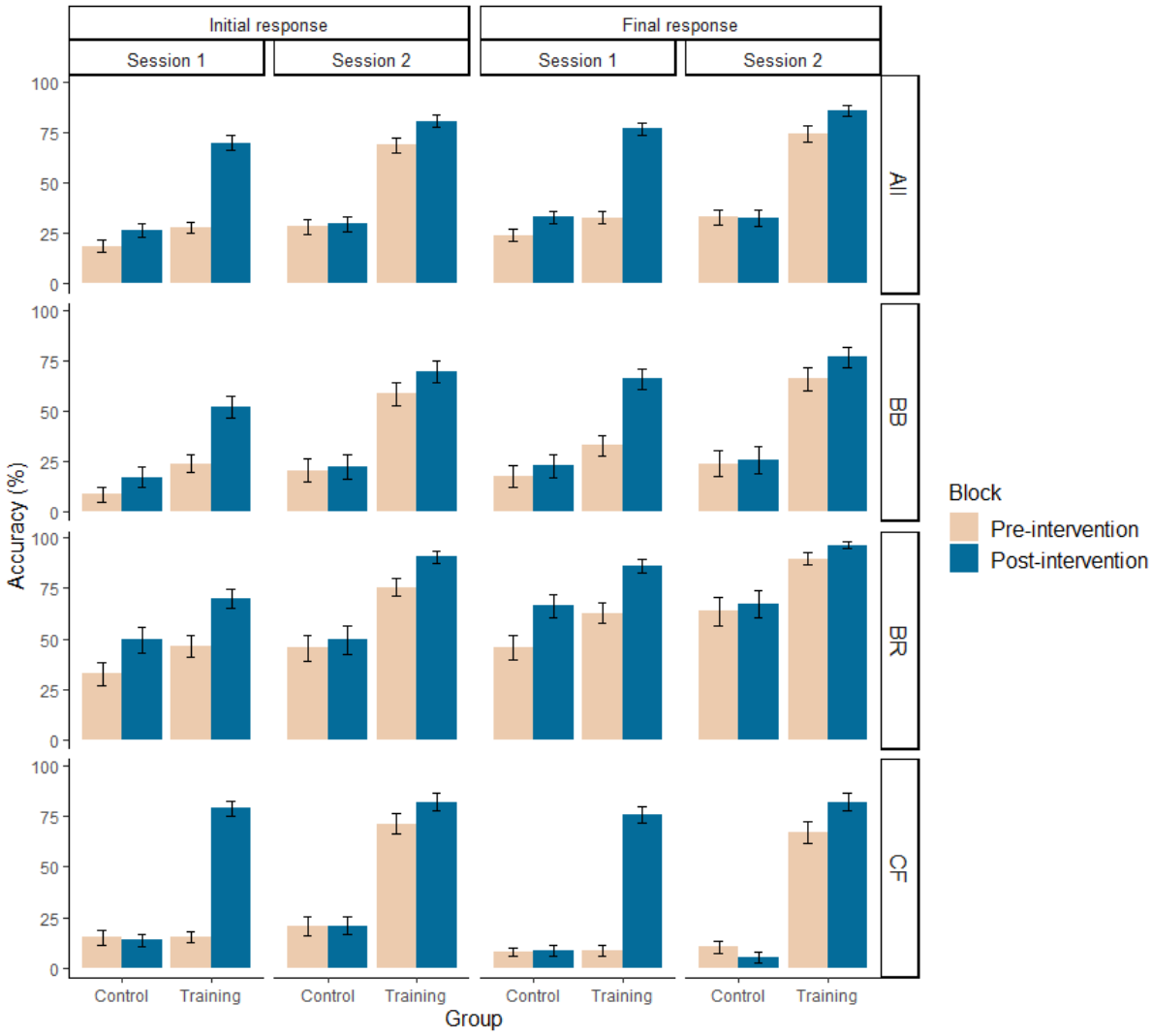


Figure 1. Mean accuracy (%) of correct initial and final responses on conflict problems for control and training groups, before and after Session 1 and Session 2, for each task (BB, BR, CF), and combined (All). Error bars are standard errors. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks, All = the composite mean across the three tasks.

Direction of change. To better understand how people changed (or did not change) their answers after deliberation, we performed a direction of change analysis for the conflict items (Bago & De Neys, 2017). Specifically, each trial is composed of two responses, the initial "intuitive" one (with time and load constraints) and the final "deliberate" one. Correct responses are labelled '1' and incorrect responses are labelled '0'. Hence, each trial can result in one of four different patterns: "00" pattern, incorrect response at both response stages; "11" pattern, correct response at both response stages; "01" pattern, initial incorrect and final correct responses; "10" pattern, initial correct and final incorrect responses. Figure 2 shows the direction of change distribution for each group in pre- and post-intervention blocks.

In line with the overall accuracies presented above, most of the time reasoners produced "00" patterns before the intervention, in both control and training groups (respectively, M = 71.3%, SD = 22.3 and M = 62.3%, SD = 24.4). However, in the training group, the intervention led to a sharp decrease in "00" patterns (43.9 points drop between pre- and post-intervention), and a considerable increase in "11" patterns (41.6 points rise). These trends were far less pronounced in the control group (respectively, a decrease of 7.9 points in "00" patterns, and a rise of 8.4 points in "11" patterns). Notably, as Boissin et al. (2022) already observed, the decrease of "00" patterns after the intervention led to an increase in "11" patterns rather than in "01" patterns (41.6 vs 2.1 points rise, following the intervention in the training group). In other words, the training helped participants intuit the correct solution strategy rather than correct an initial "erroneous" response through deliberation. Note that similar trends were observed for each of the individual reasoning tasks (see Figure S5 in Supplementary Material Section E for full results).

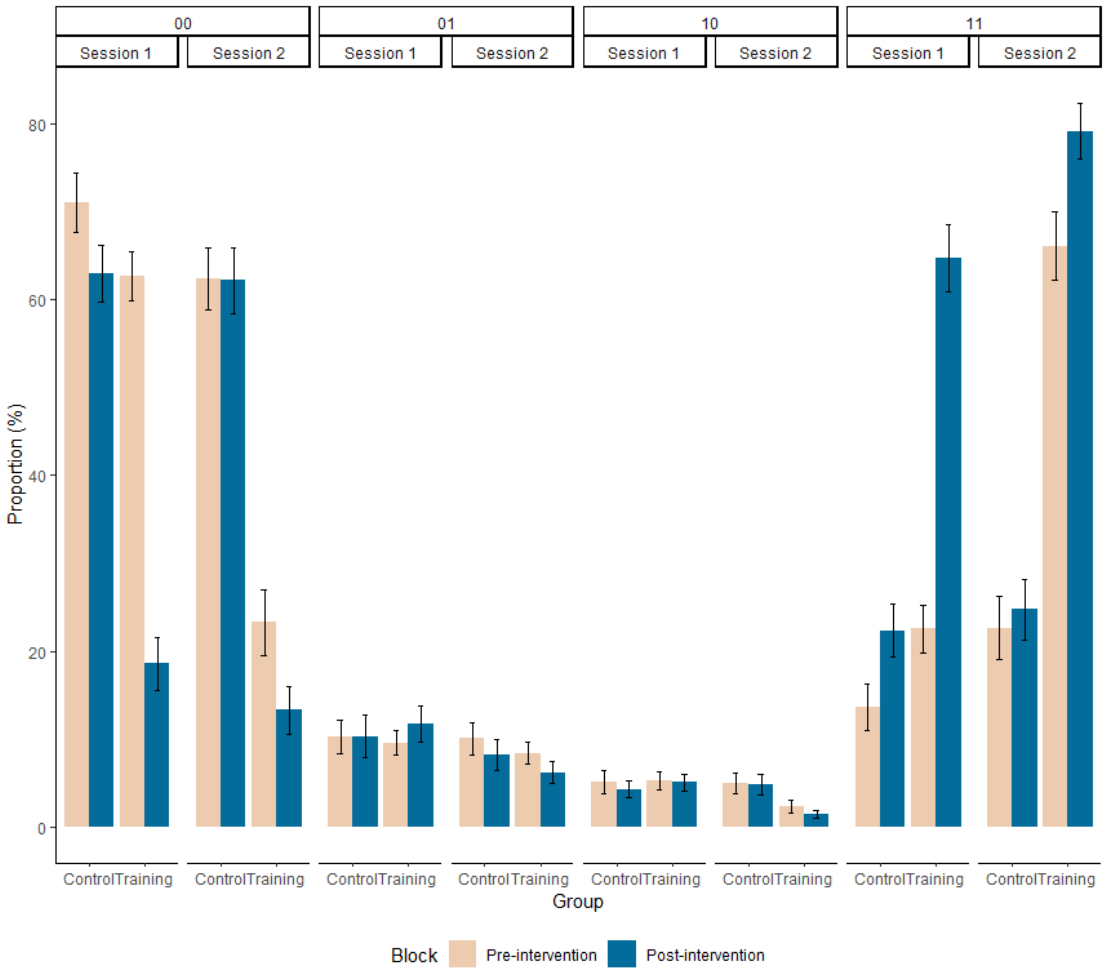


Figure 2. Proportion (%) of each direction of change (i.e., "00" trials, "01" trials, "10" trials, and "11" trials; 0 = incorrect response, 1 = correct response, first digit = initial response, second digit = final response) on conflict problems for control and training groups, before and after Session 1 and Session 2. Error bars are standard errors.

Individual level direction of change. To gain some deeper insight into how a given reasoner changed (or did not change) their response, we also performed an individual level accuracy analysis on the conflict trials (Raoelison & De Neys, 2019). For each of the 120 reasoners, we focus on their dominant direction of change and classified it using the categories introduced by Boissin et al. (2021, 2022).

First, "biased responders" did not benefit from the intervention and provided a majority of incorrect responses ("00" trials) in pre- and post-intervention blocks. They represented 66.1% of reasoners in the control group, and 23.7% in the training group. Second, "correct responders" provided a stable majority of correct answers ("01" or "11" trials) before and after the training intervention, and thus did not require any intervention to respond correctly. They represented 19.1% of reasoners in the control group and 26.5% in the training group. Third, "improved responders" are those whose accuracy increased after the training intervention. They either gave a majority of biased responses ("00" trials) before the intervention and then switched to a majority of correct responses after the intervention ("01" or "11" trials), or already gave a majority of correct final responses ("01" trials) before the intervention but then switched to a majority of correct initial and final responses ("11" trials) after the intervention. They amounted to 11.8% of reasoners in the control group and 47.9% in the training group. Participants who gave inconsistent response patterns and could not be classified were put in the "Other" category (2.9% in the control group, 1.9% in the training group; see Figure S6 in Supplementary Material Section F for full results).

No-conflict trial accuracy. As expected, the no-conflict trials analysis revealed that performance was consistently at ceiling in pre- and post-intervention blocks for initial responses (M = 84.7%, SD = 29.3 in the control group; M = 90.9%, SD = 11.6 in the training group), and for final responses (M = 88.0%, SD = 13.1 in the control group; M = 92.3%, SD = 11.6 in the training group). The high initial and final performance on the no-conflict control problems argues against a general systematic guessing confound (Bago & De Neys, 2017). It also argues against a "reversed heuristic" training account (Boissin et al., 2022) in which training would simply lead participants to distrust the intuitively cued response. This would have led to a floored post-intervention performance on the no-conflict problems (in which the intuitive, heuristic response was always correct). A detailed overview of the no-conflict problem accuracies by task can be found in Table S1 in Supplementary Material Section D.

Order effect. As we combined three reasoning tasks in a single training battery, for exploratory purposes, we also performed analyses on the presentation order. For each participant, in control and

training groups, task order was randomly assigned. Figure S9 in Supplementary Material Section G reports accuracy by task depending on the order. Note that overall, there was a trend towards a slightly less strong conflict trial training effect for the task that was presented at the end of the session. In the training group, from pre- to post-intervention, initial responses rose by 50.2 points for the first task and 34.6 points for the last task. In the same vein, final responses increased by 51.1 points for the first task and 37.3 points for the last task. This might be due to a fatigue effect near the end of the one-hour training session. However, even for the last task in the set, the training benefit over the control group was readily clear (initial responses: +34.6 points for the training group vs +9.1 points for the control group; final responses: +37.3 points for the training group vs +8.9 points for the control group).

Session 2 accuracy

The Session 1 accuracy results confirm that a single training can increase both initial and final accuracies on various classic reasoning tasks. This debiasing effect is consistent with the recent literature (e.g., see Boissin et al., 2021, 2022; Purcell et al., 2022) and shows that explaining a specific reasoning problem leads to a substantial improvement in reasoning performance, as early as the intuitive stage. After having established that our Session 1 debias findings are consistent with previous single-shot training studies, we move on to exploring the impact of the second training session. Two days after Session 1, participants were invited for a second similar one-hour training session, composed of the same three reasoning tasks (i.e., bat-and-ball, base-rate neglect, and conjunction fallacy tasks). The only difference with the first training was the material used: All the problems featured modified contents (see Supplementary Material Section A). In total, 110 participants out of 120 (i.e., 91.7%) took part in the re-test (86 females, M_{age} = 38.1 years, SD = 12.7). The sample was composed of 67 participants in the training group, and 43 in the control group. Participants from the control group in the first session again served as control group in the second session and were not given any problem explanations during the intervention.

Conflict trial accuracy. We tested whether a second training (i.e., Session 2) could further improve reasoning performance. Consequently, we compared conflict accuracies across the pre- and post-intervention blocks of Session 2. Eyeballing Figure 1 indicates that although the initial and final performance was already high in the training group after the first session, it tended to further increase after the second training. On average, compared to the pre-intervention level of Session 2, final accuracy after the second intervention further increased by 11.2 points (control group: -0.3 point), reaching M = 85.6%, SD = 22.7 in the post-intervention block. Statistical composite analyses revealed that the main final accuracy, $\chi 2$ (1) = 55.7, $p < .001$, significantly improved after the second training

and that the Block x Group interaction for final responses reached marginal significance, $\chi2$ (1) = 3.9, p = .05.

A similar second training effect was observed on the initial responses: On average, compared to the pre-intervention level of Session 2, initial accuracy after the second intervention further increased by 12.0 points (control group: +1.5 points), reaching M = 80.5%, SD = 24.9 in the post-intervention block. Statistical analyses also revealed that the main initial accuracy, $\chi2$ (1) = 50.9, p < .001 significantly improved after the second training and that the Block x Group interaction for initial responses reached marginal significance, $\chi2$ (1) = 3.5, p = .06. These trends were also observed on each individual task (see Figure 1, bottom panels).

In sum, while less pronounced than the massive first training effect (i.e., overall, approximately 11.6 points increase in Session 2 vs 43.1 points increase in Session 1), the training repetition nevertheless tended to lead to a further performance increase.

Direction of change. Figure 2 plots the conflict trial direction of change distribution, in the pre- and post-intervention blocks of Session 2. First thing to note is that the second intervention led again to a decrease in "00" patterns in the training group (from M = 23.1%, SD = 30.7 in the pre-intervention block to M = 13.5%, SD = 22.6 in the post-intervention block, i.e., 9.6 points drop), but not in the control group. Hence, regarding "00" patterns after the second intervention, we observe a gap of more than 49 points between reasoners of the control and training group. As in Session 1, the "00" patterns decrease specifically led to an increase in "11" patterns in the training group (+13.0 points), reaching a total of 79.3% of correct post-intervention responses in this group (see Figure S5 in Supplementary Material Section E for full results).

Individual level direction of change. As in Session 1, we also performed an individual level accuracy analysis, using the four categories ("correct", "biased", "improved", "other") defined by Boissin et al. (2021, 2022). Throughout Session 2, there were 70.9% of correct responders in the training group, giving a majority of "11" response patterns (vs 27.4% in the control group). We also noticed an additional training effect with an increase of 15.1% of improved reasoners, while only 4.8% of participants in the control group spontaneously improved. Finally, in the training group there only remained 12.1% biased responders (vs 66.1% in the control group, see Figure S7 in Supplementary Material Section F for full results).

Contrast between Session 1 and Session 2. For completeness, we also compared the performance in the post-intervention blocks of Session 1 and Session 2 (rather than the pre- vs post-

intervention blocks within Session 2). Not surprisingly, given that the Session 1 post-intervention and Session 2 pre-intervention performance two days later was virtually identical, results were consistent.

No conflict trial accuracy. For completeness, consistent with Session 1, no-conflict problem accuracies were also analysed. As in Session 1, performance was consistently high in pre- and post-intervention blocks for both initial and final responses (see Table S2 in Supplementary Material Section D).

Additional Session 1 and 2 analyses: Conflict detection confidence

Conflict Detection. Previous work in the reasoning field observed that biased reasoners often show some conflict or error sensitivity—as expressed for example in decreased confidence in their erroneous conflict trial responses (e.g., see De Neys, 2022, for review). As Boissin et al., (2021, 2022), we explored whether the training intervention affected biased reasoners' ability to detect conflict. That is, although the training might not have succeeded in getting all biased people to reason more accurately, it might have helped them to better detect that their answer was incorrect.

Remember that for each problem, participants were asked to enter their confidence in the correctness of their answer, on a scale from 0%, absolutely not confident, to 100%, absolutely confident (see Procedure). We used the conflict detection index introduced in the study of De Neys et al. (2011), which contrasts confidence 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 control and training groups. Hence, a higher index can be assumed to reflect a more pronounced conflict or error detection sensitivity. Following our preregistration, we focused on initial response conflict detection since it gives a purer measure of intuitively experienced conflict (e.g., see Bago & De Neys, 2017; Voudouri et al., 2022).

Boissin et al (2021, 2022) reported trends towards a better conflict detection after the training for bat-and-ball and base-rate tasks. However, regarding the conjunction fallacy task, this was not observed, and they argued against the use of the index with the specific conjunction fallacy format we adopted (see also Aczel et al., 2016; Scherer et al., 2017). We therefore analysed the results for each problem separately.

Focusing on the training group, we found a small trend towards a better conflict detection after training on bat-and-ball (from pre- to post-intervention, the index rose by 10.5 points in Session 1 and 1.2 points in Session 2) and base-rate problems (+4.6 points in Session 1 and +7.4 points in Session 2). This effect was not observed in the control group. If anything, the index tended to the opposite trend for both bat-and-ball (-1.3 points in Session 1 and +1.6 points in Session 2) and base-

rate tasks (-2.6 points in Session 1 and -5.4 points in Session 2). Hence, although some reasoners failed to provide the correct response after both bat-and-ball and base-rate training interventions, we cannot exclude that they may nevertheless have benefited from it, since they were better able to detect that their heuristic conflict problem answer was not correct after the intervention.

In line with Boissin et al. (2022), the conjunction fallacy task showed a trend towards the opposite effect in both sessions (-4.7 points in Session 1 and -0.4 point in Session 2; control group: -0.4 point in Session 1 and + 3.4 points in Session 2). For full results, see Table S4 in Supplementary Material Section H.

Predictive conflict detection. As Boissin et al. (2021, 2022), we also used confidence ratings to test the predictive effect of conflict detection, i.e., to see whether one's ability to detect conflict before the intervention could predict a better success of the training intervention. We therefore analysed whether reasoners who improved their performance after the intervention showed better conflict detection before the intervention, compared to reasoners who did not improve throughout the training (respectively, improved and biased reasoners, following the individual level direction of change classification). To calculate this predictive effect, we compared initial conflict detection of improved and biased reasoners of the training group, before the intervention, in Session 1 and Session 2.

Boissin et al (2021, 2022) reported trends towards a predictive conflict detection effect for bat-and-ball and base-rate tasks, but not for the conjunction fallacy task. In line with those results, in Session 1, for both bat-and-ball and base-rate tasks, we found a slightly better conflict detection before the training for the improved responders ($M_{improved}$ = 7.9%, SD = 18.2 for bat-and-ball; and $M_{improved}$ = 8.2%, SD = 19.0 for base-rate) compared to the biased ones (M_{biased} = 4.8%, SD = 23.3 for bat-and-ball; and M_{biased} = 4.0%, SD = 12.9 for base-rate). We did not find this predictive effect for the conjunction fallacy task, in which improved responders did not show a better conflict detection compared to the biased ones (respectively, $M_{improved}$ = 7.6%, SD = 15.7 and M_{biased} = 16.2%, SD = 16.2). Two days later, in Session 2, we found a stronger trend towards a better conflict detection before the training for the improved reasoners in both bat-and-ball and base-rate tasks (respectively, $M_{improved}$ = 36.3%, SD = 32.8 and $M_{improved}$ = 31.3%, SD = 36.9) compared to the biased reasoners (respectively, M_{biased} = 0.3%, SD = 17.3 and M_{biased} = 8.2%, SD = 11.5). This trend, although less strong, was also noticed for the conjunction fallacy task in which conflict detection of improved reasoners ($M_{improved}$ = 7.6%, SD = 21.8) was slightly higher than conflict detection of biased reasoners (M_{biased} = 4.7%, SD = 26.6). The interested reader can find details in Table S5 in Supplementary Material Section H.

In sum, in Session 1 we reproduced the trends observed by Boissin et al. (2021, 2022) on the bat-and-ball and base-rate tasks. However, in Session 2 these effects became much more pronounced. Those reasoners who remained biased after Session 1 but then improved after Session 2 were characterized by a remarkably strong conflict detection at the start of Session 2.

Study 2

Study 1 showed that a second training session further boosted reasoning performance compared to a single-shot training session. In Study 2, we aimed to test whether the training effect was robust and sustained over time. Consequently, two months after completion of Session 1, trained participants who completed Sessions 1 and 2 were invited to take part in a re-test (i.e., Study 2). Study 2 used the same procedure as Study 1. For each task, after a pre-intervention block, participants again went through our training intervention and completed a post-intervention block. This also allowed us to explore whether an additional training session (i.e., Session 3) could further boost participants' reasoning performance.

Method

Preregistration and data availability

The study design and research question were preregistered on the AsPredicted website (https://aspredicted.org) and stored on the Open Science Framework. No specific analyses were preregistered. All data and analysis scripts are available on the Open Science Framework (http://osf.io/3aqh4).

Participants

All 67 participants from the training group who completed the first two training sessions were contacted again and invited to participate. In total, 50 of them took part in Session 3 (i.e., 75%; 36 females, M_{age} = 40.6 years, SD = 14.2). We compensated participants for their time at the rate of £7 per hour.

Note that there was no control group in Session 3. For ethical reasons, control group participants were given the training explanations at the end of Session 2. Consequently, they could no longer serve as a no-training control group.

Materials and procedure

The material and the procedure were the same as in Session 1 (see Supplementary Material Section A; see Table S10 in Supplementary Material Section I for justification data).

Trial exclusion

Following our preregistration, we discarded trials in which participants failed to provide their initial answer before the deadline (1.4%) or failed to pick the correct matrix in the load task (10.0% of the remaining trials), and we analysed the remaining 90.0% of all trials. On average, each participant contributed 42.6 (SD = 4.5) conflict trials out of 48 and 42.4 (SD = 5.1) no-conflict trials out of 48.

Results

The sustained training effect

To test whether the training effect sustained over time, we compared performance on conflict items for the training group between the post-intervention block of Session 2 (i.e., after the second training) and the pre-intervention block of Session 3 (i.e., two months later).

Conflict trial accuracy. Figure 3 shows that overall performance slightly decreased after two months (final responses: M = 71.3%, SD = 31.6 in the pre-intervention block of Session 3, which corresponds to a drop of 14.3 points compared to the post-intervention block of Session 2, $t(88) = 2.7$, $p = .009$, d = 0.5; initial responses: M = 68.9%, SD = 30.6 in the pre-intervention block of Session 3, which corresponds to a drop of 11.6 points compared to the post-intervention block of Session 2, $t(91) = 2.3$, $p = .02$, d = 0.4), but reasoners still predominantly gave correct initial and final responses. This suggests that the training effect sustained over time. Importantly, note also that despite a slight decrease two months after the second session, performance in the pre-intervention block of Session 3 remains equivalent to that obtained after the first training in Session 1 (i.e., in the post-intervention block of Session 1, final responses: M = 76.9%, SD = 26.6, $t(93) = 1$, $p = .31$, d = 0.2; initial responses: M = 69.8%, SD = 31.0, $t(106) = 0.2$, $p = .88$, d = 0.03) which further indicates that the training effect is robust. These results were also backed up by a direction of change analysis (see Figure S5 in Supplementary Material Section E).

Note that in the previous single session debiasing studies of Boissin et al. (2021, 2022) with the same bat-and-ball, base-rate, and conjunction fallacy tasks, the delayed performance after two months consistently fell below that obtained after the first (i.e., single) training. In an exploratory analysis we contrasted the performance two months after the last intervention in the single training session in Boissin et al. (2021, 2022) studies and the current delayed two months performance. Results showed that the performance after two months was considerably better after the current repeated training than after the single training session (initial trial accuracy +19.2%, final trial accuracy +15.3%; see

Supplementary Material Section J for details). This tentatively suggests that repeated training led to a more enduring long-term performance.

In Session 3, we managed to reach 75% (50/67) of the Session 2 trained participants. To check for a possible attrition confound (e.g., subjects who did better in Session 2 were more likely to sign-up for Session 3), we compared the Session 2 pre-intervention conflict problem accuracy of the subgroup of Session 3 participants (initial responses: M = 69.7%, SD = 31.2; final responses: M = 75.7%, SD = 31.2) to the accuracy of Session 2 pre-intervention of the participants who did not take part (but were invited) to the re-test (initial responses: M = 65.4%, SD = 33.7; final responses: M = 70.3%, SD = 30.4). Given that both groups showed very similar accuracy rates (initial responses: $t(26) = 0.5$, $p = .65$, d = 0.1; final responses: $t(28) = 0.6$, $p = .53$, d = 0.2), it is unlikely that the Session 3 results are artificially boosted because of an attrition confound.

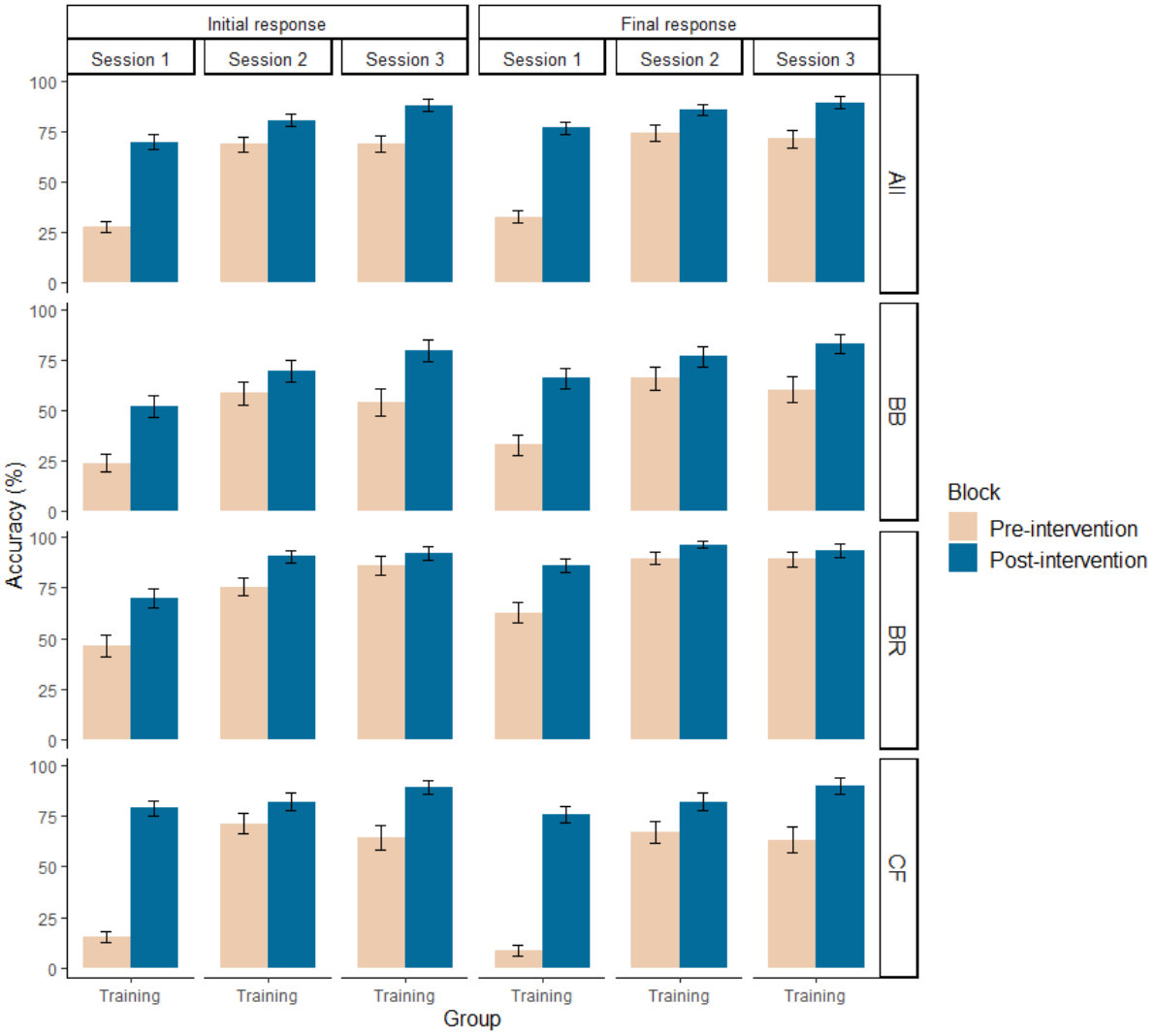


Figure 3. Mean accuracy (%) of correct initial and final responses on conflict problems for the training group, before and after each session (Session 1, Session 2, and Session 3), for each task (BB, BR, CF), and combined (All).

Error bars are standard errors. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks, All = the composite mean across the three tasks.

Third training effect

We also tested whether a third training (i.e., Session 3) could further improve reasoning performance. Consequently, we compared conflict accuracies across the pre- and post-intervention blocks of Session 3, and across the post-intervention blocks of Session 2 and of Session 3.

Conflict trial accuracy. When contrasting the Session 3 pre- and post-intervention increase, it is clear that training again boosts performance, both for final responses (+17.9 points, reaching M = 89.2%, SD = 21.5 in the post-intervention block) and initial responses (+18.9 points, reaching M = 87.8%, SD = 21.3 in the post-intervention block). Statistical composite analyses also revealed that the main final accuracy, $\chi 2$ (1) = 26.8, $p < .001$, and the main initial accuracy, $\chi 2$ (1) = 28.1, $p < .001$, significantly improved after the third training. Interestingly, although performance after the second training was already high, we found a better post-intervention Session 3 accuracy than immediately after the second training (i.e., +7.3 points initial response: $t(113) = 1.7$, $p = .09$, d = 0.3; +3.6 points final response: $t(108) = 0.9$, $p = .39$, d = 0.2). Hence, the third training seems to lead to an additional improvement. This result was also backed up by a direction of change analysis (see Figure S5 in Supplementary Material Section E).

Individual level direction of change. We performed an individual level accuracy analysis according to the type of respondent classification in Study 1 (Boissin et al., 2021, 2022). Mirroring the overall accuracy effects, a majority of reasoners were already labelled as correct (overall 68.9%) at the start of Session 3, but we still observed some improvement with the third training intervention (overall 21.0% improved reasoners). Few reasoners stayed biased and did not benefit from the third training session (overall 8.1% biased reasoners). Details can be found in Figure S8 in Supplementary Material Section F.

No conflict trial accuracy. For completeness, consistent with Study 1, no-conflict problem accuracies were also analysed. As in Study 1, performance was consistently at ceiling in pre- and post-intervention blocks for initial and final responses (see Table S3 in Supplementary Material Section D).

Additional Session 3 analyses: Conflict detection confidence. For completeness, we also looked at the conflict detection, and predictive conflict detection in Study 2. Findings were consistent with those of Study 1 for the bat-and-ball and base-rate tasks. Notably, throughout Session 3, there

was a trend towards a better conflict detection on the bat-and-ball task (the index rose by more than 10 points, going from M = 16.1%, SD = 28.4, in the pre-intervention block to M = 26.9%, SD = 33.8, in the post-intervention block of Session 3) and base-rate task (+10.4 points, going from M = 4.0%, SD = 8.0 to M = 14.6%, SD = 33.8). Regarding predictive conflict detection, improved reasoners who benefited from the training intervention showed a more pronounced conflict detection effect in the Session 3 pre-intervention block than those whose performance did not improve after a third intervention for both bat-and-ball ($M_{improved}$ = 23.5%, SD = 23.8; M_{biased} = 11.6%, SD = 20.4) and base-rate tasks ($M_{improved}$ = 20.0%, SD = 26.2; M_{biased} = 7.0%, SD = 9.6).

Interestingly, after a third training session, similar effects were also observed for the conjunction fallacy task. The conflict detection index sharply rose from M = 9.4%, SD = 15.9, in the pre-intervention block to M = 21.8%, SD = 30.9, in the post-intervention block of Session 3, and predictive conflict detection was also much higher for improved reasoners ($M_{improved}$ = 29.2%, SD = 42.9) compared to biased ones (M_{biased} = 5.3%, SD = 13.3).[2]

Hence, as in Study 1, reasoners who started to respond correctly after the training seem to be characterized by more pronounced conflict detection before the training. In other words, it seems that whereas the previous training did not yet suffice to get them to answer correctly, it did specifically boost their error detection which then served as a precursor for the intervention effect.

General Discussion

In this study, we explored whether repeating a short explanation debiasing approach on a battery of three reasoning tasks (i.e., bat-and-ball, base-rate neglect, and conjunction fallacy) can boost correct intuitive and deliberate reasoning performance. We ran three debiasing training sessions: A repetition within the same week (i.e., Session 1 and Session 2) followed by a third session two months after the initial session (i.e., Session 3). We used a two-response paradigm to track participants' initial "intuitive" and final "deliberate" responses.

Consistent with previous debiasing findings (e.g., Boissin et al., 2021, 2022), Session 1 results showed a clear first training effect (overall +41% performance increase). Across the different tasks, our short, plain-English explanation debiasing approach helped reasoners to favour the correct response over a conflicting cued heuristic mathematical response (for bat-and-ball) or a biasing stereotypical belief (for base-rate and conjunction fallacy). We observed that once the problem has been properly explained, many initially biased reasoners manage to produce correct responses to structurally similar problems afterwards. Importantly, the two-response findings indicated that this training effect was

[2] This might tentatively indicate that the format related confidence measurement confusion on the conjunction fallacy task (e.g., Aczel et al., 2016; Scherer et al., 2017) is resolved with repeated training. However, we remain to interpret the conjunction fallacy confidence findings with caution.

observed as early as the initial "intuitive" response stage. Hence, after training, reasoners can respond correctly without further need for deliberation. This validates previous findings and establishes that debiasing training can lead to sound intuiting (see Boissin et al., 2021, 2022).

However, the Session 2 (overall +11.6% performance increase) and 3 (overall further +5.4% increase) results showed that the training effect could be further boosted by repeating the training. Indeed, at the end of Session 3, the critical conflict trial accuracy approached 90% both for initial and final responses. In addition, our individual level classification indicated that at the end of Session 3 only a mere 8% of the trained group remained predominantly giving biased responses (vs. 24% after the first session). This implies that with repeated training one can virtually eliminate the infamous biased intuiting in reasoning tasks. Bluntly put, repeated debias training is not only efficient at improving the performance on some trials or for some participants, but can help almost the full sample to benefit.

Critically, Study 2 showed that the training improvements were robust and persisted after two months. Although there was a slight performance decrease at the start of the third training session, trained reasoners were still performing at the post-intervention level of the first training (and obviously above the untrained level). As we noted, previous single session debiasing studies typically observed that the delayed performance after two months fell below that obtained after the first (i.e., single) training (e.g., Boissin et al., 2021, 2022). Exploratory contrasting the delayed two months performance after these single training studies and the repeated training in the current study indicated that repeated training was associated with a better performance two months later. This tentatively suggests that repeated training also leads to a more enduring long-term performance. These results are consistent with the wider general literature on repeated testing (e.g., Higham et al., 2022; Rawson & Dunlosky, 2022) which also indicates that repetition can lead to better long-term retention.

Finally, our additional confidence analyses also allowed us to look at reasoner's conflict or error detection sensitivity. That is, even if (single) training might not have succeeded in getting all biased people to reason more accurately, it might have helped them to better detect that their answer was incorrect. In line with previous findings (Boissin et al., 2021, 2022), overall, biased reasoners tended to show an increased response doubt when they erred on conflict problems after the training. This doubt was also more pronounced pre-intervention among those reasoners who became more accurate after the intervention. Interestingly, these effects seemed to become more pronounced with repeated training. As we noted, it seems that in case the previous training did not yet suffice to get a biased reasoner to answer correctly, it did specifically boost error detection among those reasoners who started to respond correctly after the subsequent training. Hence, the increased doubt or error sensitivity may serve as a precursor for the intervention effect. In general, this underscores the role of metacognitive monitoring processes in reasoning (Ackerman & Thompson, 2017; Carpenter et al.,

2022; De Neys, 2022; Pennycook et al., 2015a) and indicates that (repeated) training may also be effective at this level.

We believe that the present work can serve as a proof-of-principle for the repeated debias training approach. At the same time, it is also clear that the approach will need to be further validated and finetuned. Hence there are a number of limitations that one needs to take in mind.

First, one possible critique is that the impact of multiple training sessions could be perceived as obvious (e.g., "the more you do something, the better you get at it"). However, it is important to consider this point in the context of reasoning and decision-making research. Despite recent debiasing successes (e.g., Claidière et al., 2017; Morewedge et al., 2015), there is a long history of failed attempts and even when successful, the resulting effects have often been relatively modest (e.g., Evans et al., 1994; Fischhoff, 1982). This has sometimes led to skepticism among cognitive scientists regarding the potential of debiasing training (Morewedge et al., 2015). Empirically demonstrating that a debiasing training works, impacts people's intuitive reasoning, and—with repeated practice—does so for virtually all individuals in a sample is anything but trivial in this respect. In addition, as we discussed above, our results also suggest that repeated training leads to a more robust long-term improvement and can be especially helpful to boost metacognitive monitoring (i.e., error detection) processes.

Second, our debias work focused on elementary logical principles in classic reasoning tasks. Clearly, these lab-based tasks remain somewhat artificial (e.g., Janssen et al., 2021; Politzer et al., 2017; Prado et al., 2020). People's erroneous personal beliefs in other contexts (e.g., climate change, conspiracy theories, or extreme political ideologies) might be more resistant to change. The generalizability of the current results to these situations or tasks clearly remains to be tested. At the same time, mastering the core underlying principles we focus on is not trivial. They remain critical for sound reasoning in a wide range of situations. For example, base-rate neglect was a key driver of the mistaken belief that Covid-19 vaccines were ineffective because most hospitalized people were vaccinated (i.e., neglecting that there were far more vaccinated than unvaccinated people in the population to start with, e.g., De Neys, 2022). Hence, we believe it is critical to attest the trainability of core logical principles in classic reasoning tasks. Nevertheless, we readily acknowledge that testing the further generalizability of the current findings remains important.

Third, one may note that in some of our Study 1 tasks, participants from the training group tended to have a better pre-intervention performance compared to participants from the control group (see Figure 1). Participants were fully randomly allocated to groups and these pre-intervention differences presumably result from random chance. Although such pre-intervention differences are not optimal, they should not affect our conclusions given that our primary focus lies in investigating the interaction effect (i.e., whether the training gain differs across groups) and there were no ceiling

effects (i.e., the training group task with the highest pre-intervention accuracy had still sufficient room to improve). Furthermore, we observed the strongest training effect on the task (i.e., conjunction fallacy) on which the pre-intervention performance was most similar. This indicates that the overall stronger training effect in the training (vs control) group was not driven by any specific confound in the training group participants (e.g., higher cognitive capacity, motivation, etc.).

Fourth, with repeated training, we managed to debias the vast majority of participants. However, even after three training sessions, some reasoners remain biased. These were characterized by low error detection. In other words, conflict detection serves as a precursor to the intervention effect. Ideally, future studies should also investigate whether this is related to more general factors, such as motivation or thinking disposition (e.g., Stanovich, 2011). It could shed light on the underlying cognitive mechanisms that may account for individual differences in bias susceptibility and the efficiency of debias interventions.

Finally, for practical reasons our first two training sessions were separated by several days and the third training session followed two months after the second. Obviously, one could try to boost the training efficacy further with more immediate and/or frequent re-training. The optimal schedule remains to be explored here. However, the current findings clearly illustrate the potential of repeated debias training. This should motivate the field to start exploring repeated debias training more seriously.

Indeed, it has generally been noted that there is a long tradition in psychological learning or intervention studies to focus on “single-session” training and that moving to multiple practice sessions holds great potential to boost training efficacy (Carpenter et al., 2022; Rawson & Dunlosky, 2022). While most of this research has focused on learning and memory performance per se (e.g., Higham et al., 2022; Rawson & Dunlosky, 2022), the current study points to a similar conclusion for debias interventions. Simply repeating the training can boost the results and allows us to help individuals who would have remained biased with one, single training. We believe this is especially important since the repeated debias approach comes at a minimal cost: It is short (the actual debias explanations take less than 5 minutes per task), does not require any intervention from a human trainer, and should be easily scalable. Given the dramatic impact and societal costs associated with biased thinking (Courbet et al., 2022; Kahneman, 2011; Milkman et al., 2009), the repeated debias approach should hold great promise here and merits to be more widely applied by scholars and practitioners.

Research data availability

Raw data, analysis scripts, and pre-registrations for these studies can be downloaded from our OSF page (http://osf.io/3aqh4).

Acknowledgements

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

References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in cognitive sciences, 21(8)*, 607-617. https://doi.org/10.1016/j.tics.2017.05.004

Aczel, B., Szollosi, A., & Bago, B. (2016). Lax monitoring versus logical intuition: The determinants of confidence in conjunction fallacy. *Thinking and Reasoning*, *22*(1), 99–117. https://doi.org/10.1080/13546783.2015.1062801

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

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

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

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

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*(3), 228–247. https://doi.org/10.1016/j.cogpsych.2010.05.004

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

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

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology*, *1*(9), 496–511. https://doi.org/10.1038/s44159-022-00089-1

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

Courbet, O., Daviot, Q., Kalamarides, V., Habib, M., Castillo, M. C., & Villemonteix, T. (2022). Promoting psychological well-being in preschool children: study protocol for a randomized controlled trial of a mindfulness-and yoga-based socio-emotional learning intervention. *Trials, 23(1)*, 1-20. https://doi.org/10.1186/s13063-022-06979-2

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

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

De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 1-68. https://doi.org/10.1017/S0140525X2200142X

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, *6*(1), e15954. https://doi.org/10.1371/journal.pone.0015954

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

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

Evans, J. S. B. (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. St. B. T., 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

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

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge, United Kingdom: Cambridge University Press.

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

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

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14(3)*, 513-525. https://doi.org/10.1037/0096-1523.14.3.513

Habib, M., & Cassotti, M. (2015). Le temps des regrets : comment le développement du regret influence-t-il la prise de décision à risque des enfants et des adolescents?. *L'Année psychologique, 115(4)*, 637-664. https://doi.org/10.3917/anpsy.154.0637

Higham, P. A., Zengel, B., Bartlett, L., & Hadwin, J. A. (2022). The Benefits of Successive Relearning on Multiple Learning Outcomes. *Journal of Educational Psychology*, *114*(5), 928-944. https://doi.org/10.1037/edu0000693

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

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

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

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

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

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced Retrieval: Absolute Spacing Enhances Learning Regardless of Relative Spacing. *Journal of Experimental Psychology: Learning Memory and Cognition*, *37*(5), 1250–1257. https://doi.org/10.1037/a0023436

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

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

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

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

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing Decisions: Improved Decision Making With a Single Training Intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129–140. https://doi.org/10.1177/2372732215600886

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015a). 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

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015b). Everyday Consequences of Analytic Thinking. *Current Directions in Psychological Science*, *24*(6), 425–432. https://doi.org/10.1177/0963721415604610

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

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

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

Purcell, Z. A., Howarth, S., Wastell, C. A., Roberts, A. J., & Sweller, N. (2022). Eye tracking and the cognitive reflection test: Evidence for intuitive correct responding and uncertain heuristic responding. *Memory & Cognition*, *50*, 348–365. https://doi.org/10.3758/s13421-021-01224-8

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

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

Rawson, K. A., & Dunlosky, J. (2022). Successive Relearning: An Underexplored but Potent Technique for Obtaining and Maintaining Knowledge. *Current Directions in Psychological Science*, *31*(4), 362–368. https://doi.org/10.1177/09637214221100484

Reyna, V. F., Weldon, R. B., & McCormick, M. (2015). Educating Intuition: Reducing Risky Decisions Using Fuzzy-Trace Theory. *Current Directions in Psychological Science*, *24*(5), 392–398. https://doi.org/10.1177/0963721415588081

Scherer, L. D., Yates, J. F., Baker, S. G., & Valentine, K. D. (2017). The Influence of Effortful Thought and Cognitive Proficiencies on the Conjunction Fallacy: Implications for Dual-Process Theories of Reasoning and Judgment. *Personality and Social Psychology Bulletin*, *43*(6), 874–887. https://doi.org/10.1177/0146167217700607

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

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

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

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

Voudouri, A., Białek, M., Domurat, A., Kowal, M., & De Neys, W. (2022). Conflict detection predicts the temporal stability of intuitive and deliberate reasoning. *Thinking & Reasoning*, 1–29. https://doi.org/10.1080/13546783.2022.2077439

Supplementary Material

A. Material: Items used in Study 1 (Session 1, Session 2) and Study 2 (Session 3)

BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks.

Items used in Session 1 (Study 1) and Session 3 (Study 2):

	Task	Conflict version	No-conflict version
1	BB	In a company there are 150 men and women in total. There are 100 more men than women. How many women are there?	In a company there are 330 men and women in total. There are 300 men. How many women are there in this company?
2	BB	A music store has 210 saxophones and flutes in total. There are 200 more saxophones than flutes. How many flutes are there?	A music store has 270 saxophones and flutes in total. There are 200 saxophones. How many flutes are there in this store?
3	BB	In a store one can choose between 320 tomatoes and avocados. There are 300 more tomatoes than avocados. How many avocados are there?	In a store one can choose between 160 tomatoes and avocados. There are 100 tomatoes. How many avocados are there in the store?
4	BB	In a kitchen there are 260 knives and spoons in total. There are 200 more knives than spoons. How many spoons are there?	In a kitchen there are 220 knives and spoons in total. There are 200 knives. How many spoons are there in the kitchen?
5	BB	A national park has 650 roses and lotus flowers in total. There are 600 more roses than lotus flowers. How many lotus flowers are there?	A national park has 380 roses and lotus flowers in total. There are 300 roses. How many lotus flowers are there in this park?
6	BB	In a stadium there are 540 volleyball and basketball players. There are 500 more volleyball players than basketball players. How many basketball players are there?	In a stadium there are 490 volleyball and basketball players. There are 400 volleyball players. How many basketball players are there in the stadium?
7	BB	A city has acquired 430 buses and trains in total. There are 400 more buses than trains. How many trains are there?	A city has acquired 610 buses and trains in total. There are 600 buses. How many trains are there in this city?
8	BB	In a store there are 480 nails and hammers in total. There are 400 more nails than hammers. How many hammers are there?	In a store there are 550 nails and hammers in total. There are 500 nails. How many hammers are there in this store?
9	BB	In a restaurant, clients have been using 250 forks and napkins. There are 200 more forks than napkins. How many napkins are there?	In a restaurant, clients have been using 230 forks and napkins. There are 200 forks. How many napkins are there in the restaurant?
10	BB	A retail clerk has to sort 280 oranges and lemons in total. There are 200 more oranges than lemons. How many lemons are there?	A retail clerk has to sort 180 oranges and lemons in total. There are 100 oranges. How many lemons are there?
11	BB	A store manager has bought 310 bananas and kiwis in total. There are 300 more bananas than kiwis. How many kiwis are there?	A store manager has bought 170 bananas and kiwis in total. There are 100 bananas. How many kiwis are there in his store?
12	BB	A store is showcasing 190 pianos and xylophones in total. There are 100 more pianos than xylophones. How many xylophones are there?	A store is showcasing 280 pianos and xylophones in total. There are 200 pianos. How many xylophones are there in this store?
13	BB	On the shelves one can find 470 screws and screwdrivers. There are 400 more screws than screwdrivers. How many screwdrivers are there?	On the shelves one can find 560 screws and screwdrivers. There are 500 screws. How many screwdrivers are there on the shelves?
14	BB	For a sports event, organizers have invited 530 players and coaches. There are 500 more players than coaches.	For a sports event, organizers have invited 510 players and coaches. There are 500 players.

		How many coaches are there?	How many coaches are there in this event?
15	BB	In a forest there are 640 mango trees and guava trees. There are 600 more mango trees than guava trees. How many mango trees are there?	In a forest there are 390 mango trees and guava trees. There are 300 mango trees. How many guava trees are there in the forest?
16	BB	In a park there are 140 adults and children in total. There are 100 more adults than children. How many children are there?	In a park there are 340 adults and children in total. There are 300 adults. How many children are there in the park?
17	BR	This study contains high school students and librarians. Person 'M' is loud. There are 5 high school students and 995 librarians. Is Person 'M' more likely to be: - A high school student? - A librarian?	This study contains high school students and librarians. Person 'M' is loud. There are 995 high school students and 5 librarians. Is Person 'M' more likely to be: - A high school student? - A librarian?
18	BR	This study contains clowns and accountants. Person 'L' is funny. There are 5 clowns and 995 accountants. Is Person 'L' more likely to be: - A clown? - An accountant?	This study contains clowns and accountants. Person 'L' is funny. There are 995 clowns and 5 accountants. Is Person 'L' more likely to be: - A clown? - An accountant?
19	BR	This study contains lab technicians and aerobics instructors. Person 'D' is active. There are 996 lab technicians and 4 aerobics instructors. Is Person 'D' more likely to be: - A lab technician? - An aerobics instructor?	This study contains lab technicians and aerobics instructors. Person 'D' is active. There are 4 lab technicians and 996 aerobics instructors. Is Person 'D' more likely to be: - A lab technician? - An aerobics instructor?
20	BR	This study contains nurses and artists. Person 'S' is creative. There are 997 nurses and 3 artists. Is Person 'S' more likely to be: - A nurse? - An artist?	This study contains nurses and artists. Person 'S' is creative. There are 3 nurses and 997 artists. Is Person 'S' more likely to be: - A nurse? - An artist?
21	BR	This study contains lawyers and gardeners. Person 'W' is argumentative. There are 3 lawyers and 997 gardeners. Is Person 'W' more likely to be: - A lawyer? - - A gardener?	This study contains lawyers and gardeners. Person 'W' is argumentative. There are 997 lawyers and 3 gardeners. Is Person 'W' more likely to be: - A lawyer? - A gardener?
22	BR	This study contains scientists and assistants. Person 'C' is intelligent. There are 4 scientists and 996 assistants. Is Person 'C' more likely to be: - A scientist? - An assistant?	This study contains scientists and assistants. Person 'C' is intelligent. There are 996 scientists and 4 assistants. Is Person 'C' more likely to be: - A scientist? - An assistant?
23	BR	This study contains I.T. technicians and boxers. Person 'F' is strong. There are 995 I.T. technicians and 5 boxers. Is Person 'F' more likely to be: - An I.T. technician? - A boxer?	This study contains I.T. technicians and boxers. Person 'F' is strong. There are 5 I.T. technicians and 995 boxers. Is Person 'F' more likely to be: - An I.T. technician? - A boxer?

24	BR	This study contains businessmen and firemen. Person 'K' is brave. There are 996 businessmen and 4 firemen. Is Person 'K' more likely to be: - A businessman? - A fireman?	This study contains businessmen and firemen. Person 'K' is brave. There are 4 businessmen and 996 firemen. Is Person 'K' more likely to be: - A businessman? - A fireman?
25	BR	This study contains flight attendants and surgeons. Person 'E' is kind. There are 5 flight attendants and 995 surgeons. Is Person 'E' more likely to be: - A flight attendant? - A surgeon?	This study contains flight attendants and surgeons. Person 'E' is kind. There are 995 flight attendants and 5 surgeons. Is Person 'E' more likely to be: - A flight attendant? - A surgeon?
26	BR	This study contains accountants and boys. Person 'H' is immature. There are 997 accountants and 3 boys. Is Person 'H' more likely to be: - An accountant? - A boy?	This study contains accountants and boys. Person 'H' is immature. There are 3 accountants and 997 boys. Is Person 'H' more likely to be: - An accountant? - A boy?
27	BR	This study contains consultants and construction workers. Person 'P' is helpful. There are 4 consultants and 996 construction workers. Is Person 'P' more likely to be: - A consultant? - A construction worker?	This study contains consultants and construction workers. Person 'P' is helpful. There are 996 consultants and 4 construction workers. Is Person 'P' more likely to be: - A consultant? - A construction worker?
28	BR	This study contains high school coaches and dentists. Person 'O' is loud. There are 3 high school coaches and 997 dentists. Is Person 'O' more likely to be: - A high school coach? - A dentist?	This study contains high school coaches and dentists. Person 'O' is loud. There are 997 high school coaches and 3 dentists. Is Person 'O' more likely to be: - A high school coach? - A dentist?
29	BR	This study contains rich people and gardeners. Person 'G' is arrogant. There are 4 rich people and 996 gardeners. Is Person 'G' more likely to be: - A rich person? - A gardener?	This study contains rich people and gardeners. Person 'G' is arrogant. There are 996 rich people and 4 gardeners. Is Person 'G' more likely to be: - A rich person? - A gardener?
30	BR	This study contains women and drummers. Person 'I' is loud. There are 997 women and 3 drummers. Is Person 'I' more likely to be: - A woman? - A drummer?	This study contains women and drummers. Person 'I' is loud. There are 3 women and 997 drummers. Is Person 'I' more likely to be: - A woman? - A drummer?
31	BR	This study contains real estate agents and poor people. Person 'K' is persuasive. There are 5 real estate agents and 995 poor people. Is Person 'K' more likely to be: - A real estate agent? - A poor people?	This study contains real estate agents and poor people. Person 'K' is persuasive. There are 995 real estate agents and 5 poor people. Is Person 'K' more likely to be: - A real estate agent? - A poor people?
32	BR	This study contains secretaries and telemarketers. Person 'J' is persuasive.	This study contains secretaries and telemarketers. Person 'J' is persuasive.

		There are 995 secretaries and 5 telemarketers. Is Person 'J' more likely to be: - A secretary? - A telemarketer?	There are 995 telemarketers and 5 secretaries. Is Person 'J' more likely to be: - A secretary? A telemarketer?
33	CF	Piper, 25, has previously studied aerodynamics and likes extreme sports. Is it most probable that the described person is: - A history teacher and a motorcycle racer - A history teacher - A history teacher and a scrabble player - A mortician	Allen, 45, has previously studied aerodynamics and likes extreme sports. Is it most probable that the described person is: - A mortician - A motorcycle racer - A history teacher and a scrabble player - A history teacher and a motorcycle racer
34	CF	Corey, 36, has previously studied journalism and likes gossip. Is it most probable that the described person is: - A mine-clearer - A forest ranger and a handyman - A forest ranger - A forest ranger and a tabloid reader	Aidan, 25, has previously studied journalism and likes gossip. Is it most probable that the described person is: - A mine-clearer - A tabloid reader - A forest ranger and a handyman - A forest ranger and a tabloid reader
35	CF	Perry, 36, has previously studied literature and likes poetry. Is it most probable that the described person is: - A carpenter and a hockey player - A carpenter - An Olympic medalist - A carpenter and a novel writer	Cecil, 34, has previously studied literature and likes poetry. Is it most probable that the described person is: - A carpenter and a hockey player - A novel writer - An Olympic medalist - A carpenter and a novel writer
36	CF	Maddy, 30, has previously studied gastronomy and likes French food. Is it most probable that the described person is: - A Court of Appeal Judge - A gardener and a wine taster - A gardener - A gardener and a weightlifter	Clare, 40, has previously studied gastronomy and likes French food. Is it most probable that the described person is: - A Court of Appeal Judge - A gardener and a weightlifter - A gardener and a wine taster - A wine taster
37	CF	Blake, 39, has previously studied comedy and likes laughing. Is it most probable that the described person is: - An archivist and a karateka - An archivist - A bank CEO - An archivist and a clown	Riley, 33, has previously studied comedy and likes laughing. Is it most probable that the described person is: - A clown - An archivist and a clown - A bank CEO - An archivist and a karateka
38	CF	Briar, 30, has previously studied economics and likes quality tobacco. Is it most probable that the described person is: - A shop assistant - A shop assistant and a cigar smoker - A shop assistant and a ballet dancer - A snowboard professional	Flinn, 40, has previously studied economics and likes quality tobacco. Is it most probable that the described person is: - A cigar smoker - A shop assistant and a cigar smoker - A shop assistant and a ballet dancer - A snowboard professional
39	CF	Errin, 27, has previously studied pattern design and likes sewing. Is it most probable that the described person is: - A caregiver and a fashion enthusiast - A caregiver - An astronaut - A caregiver and a genealogist	Kelly, 43, has previously studied pattern design and likes sewing. Is it most probable that the described person is: - A caregiver and a genealogist - An astronaut - A fashion enthusiast - A caregiver and a fashion enthusiast

40	CF	Edwin, 38, has previously studied astronomy and likes sci-fi. Is it most probable that the described person is: - A longshoreman - An Oscar winner - A longshoreman and an equestrian - A longshoreman and a stargazer	Kadin, 32, has previously studied astronomy and likes sci-fi. Is it most probable that the described person is: - A stargazer - An Oscar winner - A longshoreman and a stargazer - A longshoreman and an equestrian
41	CF	Falon, 26, has previously studied education and likes children. Is it most probable that the described person is: - A flight attendant - A flight attendant and a dad - A duke - A flight attendant and a rally racing fan	Logan, 44, has previously studied education and likes children. Is it most probable that the described person is: - A duke - A flight attendant and a rally racing fan - A flight attendant and a dad - A dad
42	CF	Damon, 27, has previously studied linguistics and likes storytelling. Is it most probable that the described person is: - A heavyweight boxer - A machine operator and a free climber - A machine operator - A machine operator and a book lover	Sandy, 43, has previously studied linguistics and likes storytelling. Is it most probable that the described person is: - A heavyweight boxer - A machine operator and a free climber - A book lover - A machine operator and a book lover
43	CF	Wayne, 39, has previously studied zoology and likes mountain nature. Is it most probable that the described person is: - A navy admiral - A musician and a birdwatcher - A musician - A musician and a juggler	Flynn, 31, has previously studied zoology and likes mountain nature. Is it most probable that the described person is: - A navy admiral - A musician and a birdwatcher - A birdwatcher - A musician and a juggler
44	CF	Corri, 26, has previously studied web marketing and likes social media. Is it most probable that the described person is: - A fireman - A fireman and a puzzle lover - A fireman and a youtuber - A sword swallower	Ethan, 44, has previously studied web marketing and likes social media. Is it most probable that the described person is: - A youtuber - A sword swallower - A fireman and a youtuber - A fireman and a puzzle lover
45	CF	Billy, 27, has previously studied geography and likes foreign culture. Is it most probable that the described person is: - A pawnbroker and a globetrotter - A pawnbroker and a perfumer - A pawnbroker - A globetrotter	Billy, 27, has previously studied geography and likes foreign culture. Is it most probable that the described person is: - A pawnbroker and a globetrotter - A pawnbroker and a perfumer - A swordsman - A globetrotter
46	CF	Haven, 35, has previously studied gender studies and likes hardcore music. Is it most probable that the described person is: - An Archbishop - A shoemaker and a Jeovah witness - A shoemaker - A shoemaker and a feminist	Tommy, 35, has previously studied gender studies and likes hardcore music. Is it most probable that the described person is: - A feminist - A shoemaker and a feminist - An Archbishop - A shoemaker and a Jeovah witness
47	CF	Julia, 31, has previously studied cultural analysis and likes Apple products. Is it most probable that the described person is: - A house painter and a carpet weaver - A corporal	Jodie, 39, has previously studied cultural analysis and likes Apple products. Is it most probable that the described person is: - An iPad owner - A corporal

		- A house painter and an iPad owner - A house painter	- A house painter and an iPad owner - A house painter and a carpet weaver
48	CF	Bryce, 41, has previously studied performing arts and likes sports. Is it most probable that the described person is: - A head of state - A fruit picker and an acrobat - A fruit picker and a video gamer - A fruit picker	Paige, 31, has previously studied performing arts and likes sports. Is it most probable that the described person is: - An acrobat - A head of state - A fruit picker and an acrobat - A fruit picker and a video gamer

Items used in Session 2 (Study 1):

	Task	Conflict version	No-conflict version
1	BB	In a building residents have 370 dogs and cats in total. There are 300 more dogs than cats. How many cats are there?	In a building residents have 110 dogs and cats in total. There are 100 dogs. How many cats are there in the building?
2	BB	To make yogurt, a cook has bought 270 apricots and pears. There are 200 more apricots than pears. How many pears are there?	To make yogurt, a cook has bought 210 apricots and pears. There are 200 apricots. How many pears did the cook buy?
3	BB	At a convention there are 560 neuroscientists and botanists. There are 500 more neuroscientists than botanists. How many botanists are there?	At a convention there are 470 neuroscientists and botanists. There are 400 neuroscientists. How many botanists are there in this convention?
4	BB	A woodwork company has bought 460 drills and hacksaws. There are 400 more drills than hacksaws. How many hacksaws are there?	A woodwork company has bought 570 drills and hacksaws. There are 500 drills. How many hacksaws are there in this company?
5	BB	A retail clerk has to sort 290 oranges and lemons in total. There are 200 more oranges than lemons. How many lemons are there?	A retail clerk has to sort 180 oranges and lemons in total. There are 100 oranges. How many lemons are there for him to sort?
6	BB	The kitchen in a restaurant has 240 plates and pans in total. There are 200 more plates than pans. How many pans are there?	The kitchen in a restaurant has 250 plates and pans in total. There are 200 plates. How many pans are there?
7	BB	Around a lake there are 610 daisies and jasmine flowers. There are 600 more daisies than jasmine flowers. How many jasmine flowers are there?	Around a lake there are 430 daisies and jasmine flowers. There are 400 daisies. How many jasmine flowers are there around this lake?
8	BB	In a city people use 380 scooters and bicycles in total. There are 300 more scooters than bicycles. How many bicycles are there?	In a city people use 650 scooters and bicycles in total. There are 600 scooters. How many bicycles are there in this city?
9	BB	On a safari tour one can watch 350 lions and pumas in total. There are 300 more lions than pumas. How many pumas are there?	On a safari tour one can watch 130 lions and pumas in total. There are 100 lions. How many pumas are there on the tour?
10	BB	In a school there are 350 boys and girls in total. There are 300 more boys than girls. How many girls are there in the school?	In a school there are 350 boys and girls in total. There are 300 boys. How many girls are there in the school?
11	BB	A sports facility is housing 510 football players and swimmers. There are 500 more football players than swimmers. How many swimmers are there?	A sports facility is housing 520 football players and swimmers. There are 500 football players. How many swimmers are there in this facility?
12	BB	In a city park there are 390 skateboarders and pedestrians.	In a city park there are 640 skateboarders and pedestrians.

		There are 300 more skateboarders than pedestrians. How many pedestrians are there?	There are 600 skateboarders. How many pedestrians are there in this park?
13	BB	In a grass plain scientists have counted 330 zebras and elephants. There are 300 more zebras than elephants. How many elephants are there?	In a grass plain scientists have counted 150 zebras and elephants. There are 100 zebras. How many elephants are there in this plain?
14	BB	A music school is renting 170 guitars and harps in total. There are 100 more guitars than harps. How many harps are there?	A music school is renting 310 guitars and harps in total. There are 300 guitars. How many harps are there in this school?
15	BB	In a greenhouse there are 620 dandelions and water lilies. There are 600 more dandelions than water lilies. How many water lilies are there?	In a greenhouse there are 420 dandelions and water lilies. There are 400 dandelions. How many water lilies are there in the greenhouse?
16	BB	For a convention organizers have bought 240 glasses and cups. There are 200 more glasses than cups. How many cups did the organizers buy?	For a convention organizers have bought 240 glasses and cups. There are 200 glasses. How many cups did the organizers buy?
17	BR	This study contains computer programmers and hippies. Person 'B' is unconventional. There are 5 hippies and 995 computer programmers. Is Person 'B' more likely to be: - A computer programmer? - A hippie?	This study contains computer programmers and hippies. Person 'B' is unconventional. There are 5 computer programmers and 995 hippies. Is Person 'B' more likely to be: - A hippie? - A computer programmer?
18	BR	This study contains accountants and boys. Person 'G' is organized. There 4 accountants and 996 boys. Is Person 'G' more likely to be: - An accountant? - A boy?	This study contains accountants and boys. Person 'G' is organized. There are 4 boys and 996 accountants. Is Person 'G' more likely to be: - An accountant? - A boy?
19	BR	This study contains artists and nurses. Person 'T' is helpful. There are 997 artists and 3 nurses. Is Person 'T' more likely to be: - An artist? - A nurse?	This study contains artists and nurses. Person 'T' is helpful. There are 997 nurses and 3 artists. Is Person 'T' more likely to be: - An artist? - A nurse?
20	BR	This study contains consultants and boxers. Person 'A' is strong. There are 995 consultants and 5 boxers. Is Person 'A' more likely to be: - A boxer? - A consultant?	This study contains consultants and boxers. Person 'A' is strong. There are 995 boxers and 5 consultants. Is Person 'A' more likely to be: - A consultant? - A boxer?
21	BR	This study contains architects and telemarketers. Person 'Q' is creative. There are 3 architects and 997 telemarketers. Is Person 'Q' more likely to be: - A telemarketer? - An architect?	This study contains architects and telemarketers. Person 'Q' is creative. There are 3 telemarketers and 997 architects. Is Person 'Q' more likely to be: - A telemarketer? - An architect?
22	BR	This study contains lab technicians and politicians. Person 'E' is intelligent. There are 5 lab technicians and 995 politicians. Is Person 'E' more likely to be: - A lab technician? - A politician?	This study contains lab technicians and politicians. Person 'E' is intelligent. There are 5 politicians and 995 lab technicians. Is Person 'E' more likely to be: - A lab technician? - A politician?

23	BR	This study contains rich people and paramedics. Person 'J' is reliable. There are 996 rich people and 4 paramedics. Is person 'J' more likely to be: - A rich people? - A paramedic?	This study contains rich people and paramedics. Person 'J' is reliable. There are 996 paramedics and 4 rich people. Is Person 'J' more likely to be: - A paramedic? - A rich people?
24	BR	This study contains nannies and businessmen. Person 'C' is ambitious. There are 997 nannies and 3 businessmen. Is Person 'C' more likely to be: - A nanny? - A businessman?	This study contains nannies and businessmen. Person 'C' is ambitious. There are 997 businessmen and 3 nannies. Is Person 'C' more likely to be: - A businessman? - A nanny?
25	BR	This study contains high school coaches and dentists. Person 'O' is loud. There are 3 high school coaches and 997 dentists. Is Person 'O' more likely to be: - A high school coach? - A dentist?	This study contains high school coaches and dentists. Person 'O' is loud. There are 997 high school coaches and 3 dentists. Is Person 'O' more likely to be: - A high school coach? - A dentist?
26	BR	This study contains writers and sixteen-year-olds. Person 'Z' is immature. There are 996 writers and 4 sixteen-year-olds. Is Person 'Z' more likely to be: - A writer? - A sixteen-year-old?	This study contains writers and sixteen-year-olds. Person 'Z' is immature. There are 996 sixteen-year-olds and 4 writers. Is Person 'Z' more likely to be: - A writer? - A sixteen-year-old?
27	BR	This study contains flight attendants and scientists. Person 'H' is intelligent. There are 997 flight attendants and 3 scientists. Is Person 'H' more likely to be: - A scientist? - A flight attendant?	This study contains flight attendants and scientists. Person 'H' is intelligent. There are 3 flight attendants and 997 scientists. Is Person 'H' more likely to be: - A flight attendant? - A scientist?
28	BR	This study contains clowns and dentists. Person 'R' is funny. There are 4 clowns and 996 dentists. Is Person 'R' more likely to be: - A clown? A dentist?	This study contains clowns and dentists. Person 'R' is funny. There are 996 clowns and 4 dentists. Is Person 'R' more likely to be: - A clown? - A dentist?
29	BR	This study contains I.T. technicians and real estate agents. Person 'U' is nerdy. There are 997 real estate agents and 3 I.T. technicians. Is Person 'U' more likely to be: - An I.T. technician? - A real estate agent?	This study contains I.T. technicians and real estate agents. Person 'U' is nerdy. There are 997 I.T. technicians and 3 real estate agents. Is Person 'U' more likely to be: - An I.T. technician? - A real estate agent?
30	BR	This study contains lawyers and gardeners. Person 'X' is gentle. There are 5 gardeners and 995 lawyers. Is Person 'X' more likely to be: - A gardener? - A lawyer?	This study contains lawyers and gardeners. Person 'X' is gentle. There are 5 lawyers and 995 gardeners. Is Person 'X' more likely to be: - A lawyer? - A gardener?
31	BR	This study contains women and drummers. Person 'M' is sensitive. There 4 women and 996 drummers.	This study contains women and drummers. Person 'M' is sensitive. There 4 drummers and 996 women.

		Is Person 'M' more likely to be: - A drummer? - A woman?	Is Person 'M' more likely to be: - A drummer? - A woman?
32	BR	This study contains lab technicians and aerobics instructors. Person 'D' is intelligent. There 996 aerobics instructors and 4 lab technicians. Is Person 'D' more likely to be: - An aerobics instructor? - A lab technician?	This study contains lab technicians and aerobics instructors. Person 'D' is intelligent. There 4 aerobics instructors and 996 lab technicians. Is Person 'D' more likely to be: - An aerobics instructor? - A lab technician?
33	CF	Emery, 27, has previously studied robotics and likes AI. Is it most probable that the described person is: - A cashier and a computer hacker - A cashier and a cheerleader - A cashier - An international pop singer	Alvin, 43, has previously studied robotics and likes AI. Is it most probable that the described person is: - A cashier and a cheerleader - A cashier and a computer hacker - A computer hacker - An international pop singer
34	CF	Glenn, 40, has previously studied military strategy and likes combat sports. Is it most probable that the described person is: - A paleontologist - An insurer - An insurer and a knitter - An insurer and a gun owner	Aston, 30, has previously studied military strategy and likes combat sports. Is it most probable that the described person is: - An insurer and a knitter - An insurer and a gun owner - A paleontologist - A gun owner
35	CF	Tobey, 33, has previously studied biology and likes forest excursions. Is it most probable that the described person is: - A fighter pilot - A masseur and a mushroom picker - A masseur and a wrestler - A masseur	Ariel, 37, has previously studied biology and likes forest excursions. Is it most probable that the described person is: - A masseur and a mushroom picker - A mushroom picker - A fighter pilot - A masseur and a wrestler
36	CF	Lewis, 36, has previously studied Mechanics and likes steamships. Is it most probable that the described person is: - A waiter and a blogger - A waiter - A waiter and a boat lover - An opera singer	Lenny, 34, has previously studied Mechanics and likes steamships. Is it most probable that the described person is: - A waiter and a boat lover - A waiter and a blogger - An opera singer - A boat lover
37	CF	Jamie, 42, has previously studied sea winds and likes to sail. Is it most probable that the described person is: - A postal worker - a postal worker and a car collector - A rock star - A postal worker and a fisherman	Angel, 28, has previously studied sea winds and likes to sail. Is it most probable that the described person is: - A postal worker and a car collector - A rock star - A postal worker and a fisherman - A fisherman
38	CF	Katie, 32, has previously studied fine arts and likes painting. Is it most probable that the described person is: - A brain surgeon - A parking attendant - A parking attendant and a snowboarder - A parking attendant and a cartoonist	Lexie,38, has previously studied fine arts and likes painting. Is it most probable that the described person is: - A parking attendant and a cartoonist - A parking attendant and a snowboarder - A brain surgeon - A cartoonist
39	CF	Jenny, 33, has previously studied political science and likes local politics.	Grady, 37, has previously studied political science and likes local politics.

		Is it most probable that the described person is: - A receptionist - A princess - A receptionist and a poker player - A receptionist and a political party member	Is it most probable that the described person is: - A princess - A receptionist and a political party member - A receptionist and a poker player - A political party member
40	CF	Wyatt, 42, has previously studied musicology and likes jazz. Is it most probable that the described person is: - A taxi driver and an orienteer - An ostrich farmer - A taxi driver - A taxi driver and a record collector	Brook, 28, has previously studied musicology and likes jazz. Is it most probable that the described person is: - A taxi driver and an orienteer - An ostrich farmer - A taxi driver and a record collector - A record collector
41	CF	Marin, 29, has previously studied sound engineering and likes hifi speakers. Is it most probable that the described person is: - A countess - A baker - A baker and a music lover - A baker and an extreme sportsman	Jerry,41, has previously studied sound engineering and likes hifi speakers. Is it most probable that the described person is: - A baker and a music lover - A countess - A baker and an extreme sportsman - A music lover
42	CF	Alexa, 35, has previously studied sociology and likes trade unions. Is it most probable that the described person is: - A bus driver and a social democrat - A bus driver and a stock speculator - A rock star - A bus driver	Jaden, 35, has previously studied sociology and likes trade unions. Is it most probable that the described person is: - A bus driver and a stock speculator - A bus driver and a social democrat - A social democrat - A rock star
43	CF	Aaron, 40, has previously studied handicrafts and likes pottery. Is it most probable that the described person is: - A game show winner - A tour guide - A tour guide and a sniper - A tour guide and a woodcarver	Danny, 30, has previously studied handicrafts and likes pottery. Is it most probable that the described person is: - A woodcarver - A game show winner - A tour guide and a sniper - A tour guide and a woodcarver
44	CF	Shawn, 40, has previously studied real estate and likes luxury items. Is it most probable that the described person is: - A courier - A submarine captain - A courier and a make-up artist - A courier and a watch collector	Faith, 32, has previously studied real estate and likes luxury items. Is it most probable that the described person is: - A submarine captain - A courier and a watch collector - A watch collector - A courier and a make-up artist
45	CF	Blair, 32, has previously studied theology and likes choral singing. Is it most probable that the described person is: - A warehouse worker and a Christian - A Formula 1 driver - A warehouse worker and a paintball player - A warehouse worker	Tatum, 38, has previously studied theology and likes choral singing. Is it most probable that the described person is: - A warehouse worker and a paintball player - A Christian - A Formula 1 driver - A warehouse worker and a Christian
46	CF	Chris, 31, has previously studied computer science and likes Japanese comics. Is it most probable that the described person is: - A bartender - A bartender and an online gamer - A bartender and a pipe smoker - A diplomat	Doris, 39, has previously studied computer science and likes Japanese comics. Is it most probable that the described person is: - A diplomat - An online gamer - A bartender and an online gamer - A bartender and a pipe smoker

47	CF	Amber, 28, has previously studied mathematics and likes board games. Is it most probable that the described person is: - A guard and a martial artist - A guard and a chess player - A moose farmer - A guard	Marty,33, has previously studied mathematics and likes board games. Is it most probable that the described person is: - A chess player - A moose farmer - A guard and a chess player - A guard and a martial artist
48	CF	Gavyn, 41, has previously studied marketing and likes to deceive. Is it most probable that the described person is: - An ant farmer - A bodyguard and a poker player - A bodyguard and a nature lover - A bodyguard	Umber, 39, has previously studied marketing and likes to deceive. Is it most probable that the described person is: - An ant farmer - A bodyguard and a nature lover - A bodyguard and a poker player - A poker player

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

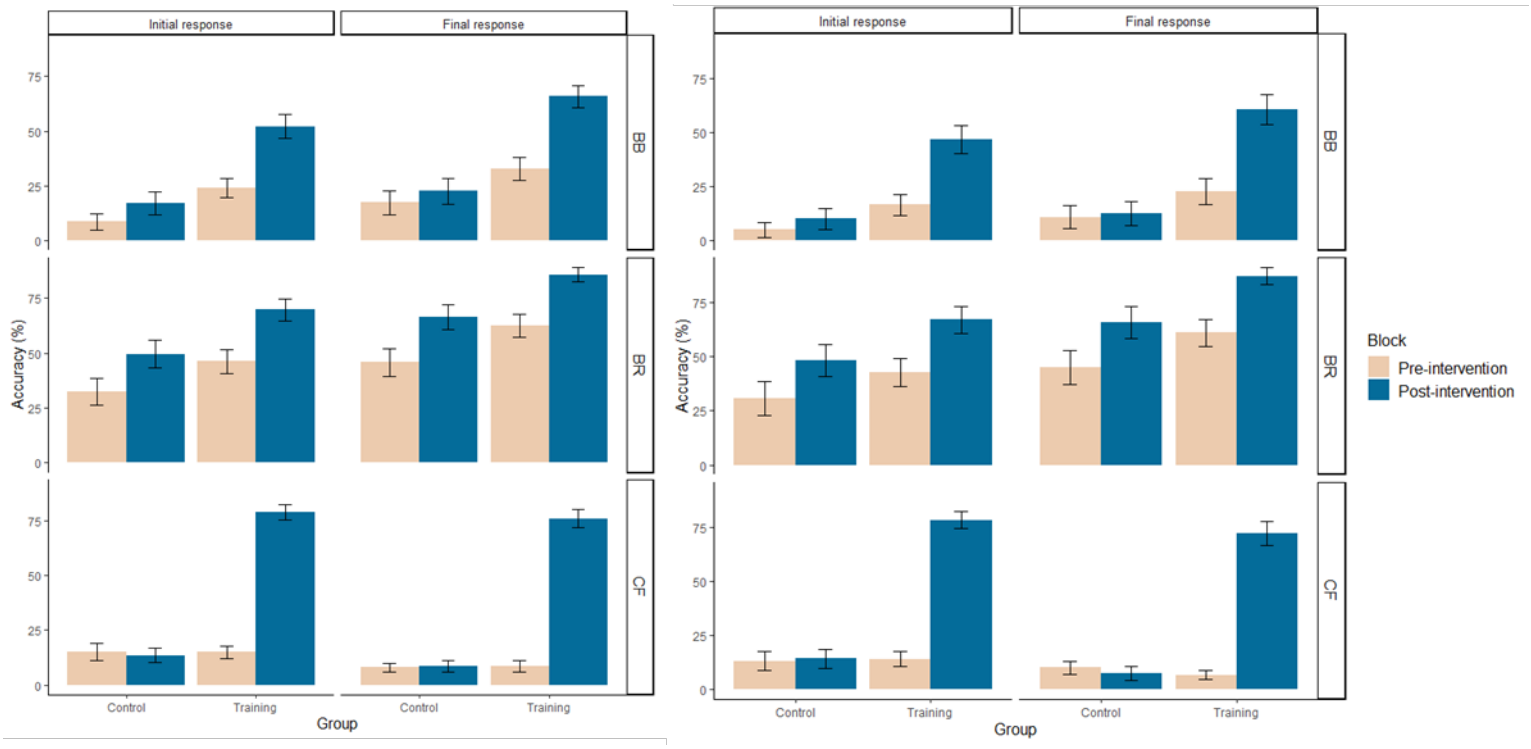


Figure S1. Mean accuracy (%) of correct initial and final responses on conflict problems before and after Session 1, with (left panel) and without (right panel) reasoners who already knew the original bat-and-ball problem (Frederick, 2005). Error bars are standard errors. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks.

C. Conjunction fallacy problems: Frequency of each individual response option in Session 1, Session 2 and Session 3

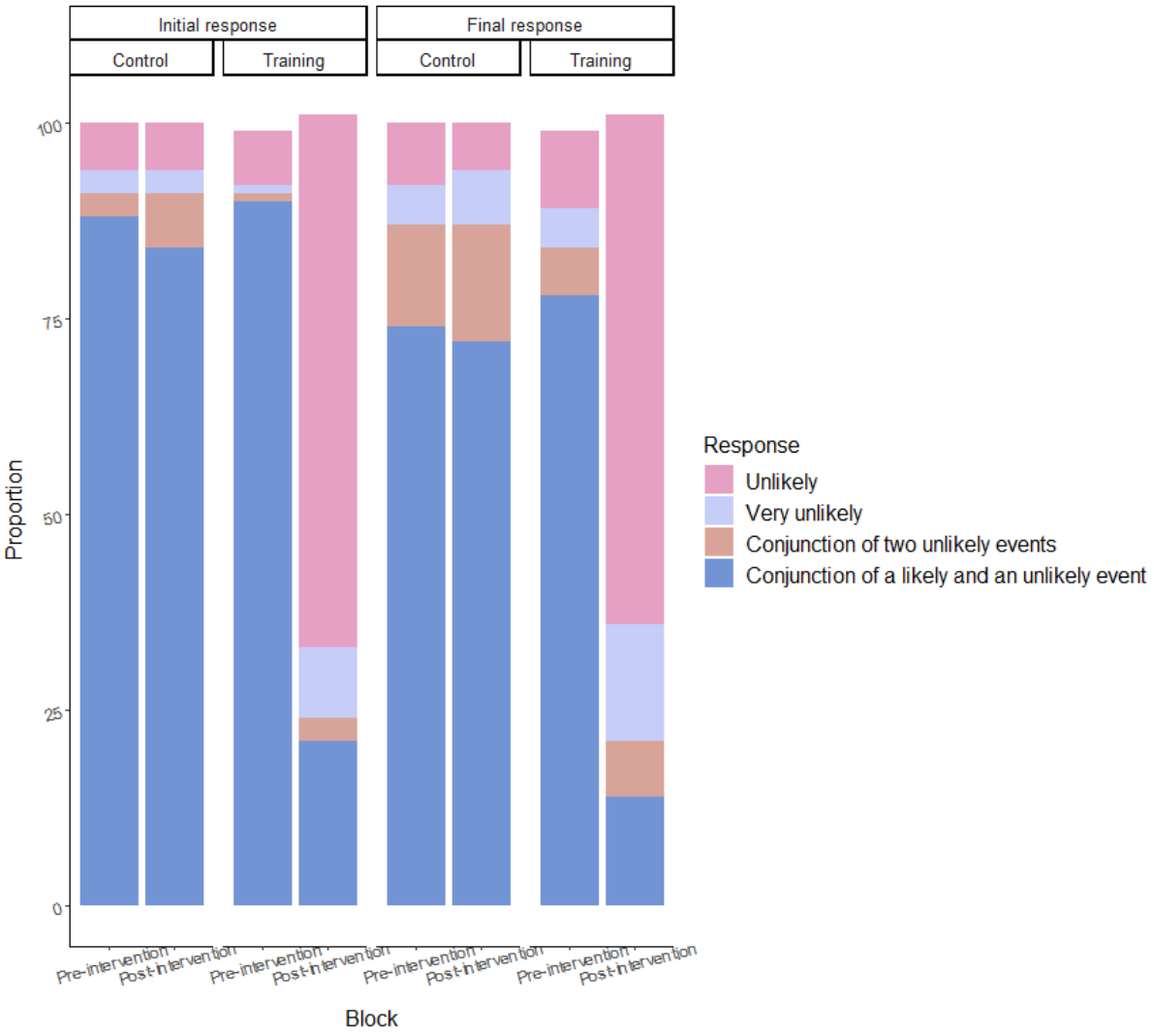


Figure S2. Frequency of each individual response option in Session 1 (conjunction fallacy items) for the initial and the final responses, before and after the intervention in the control and training group.

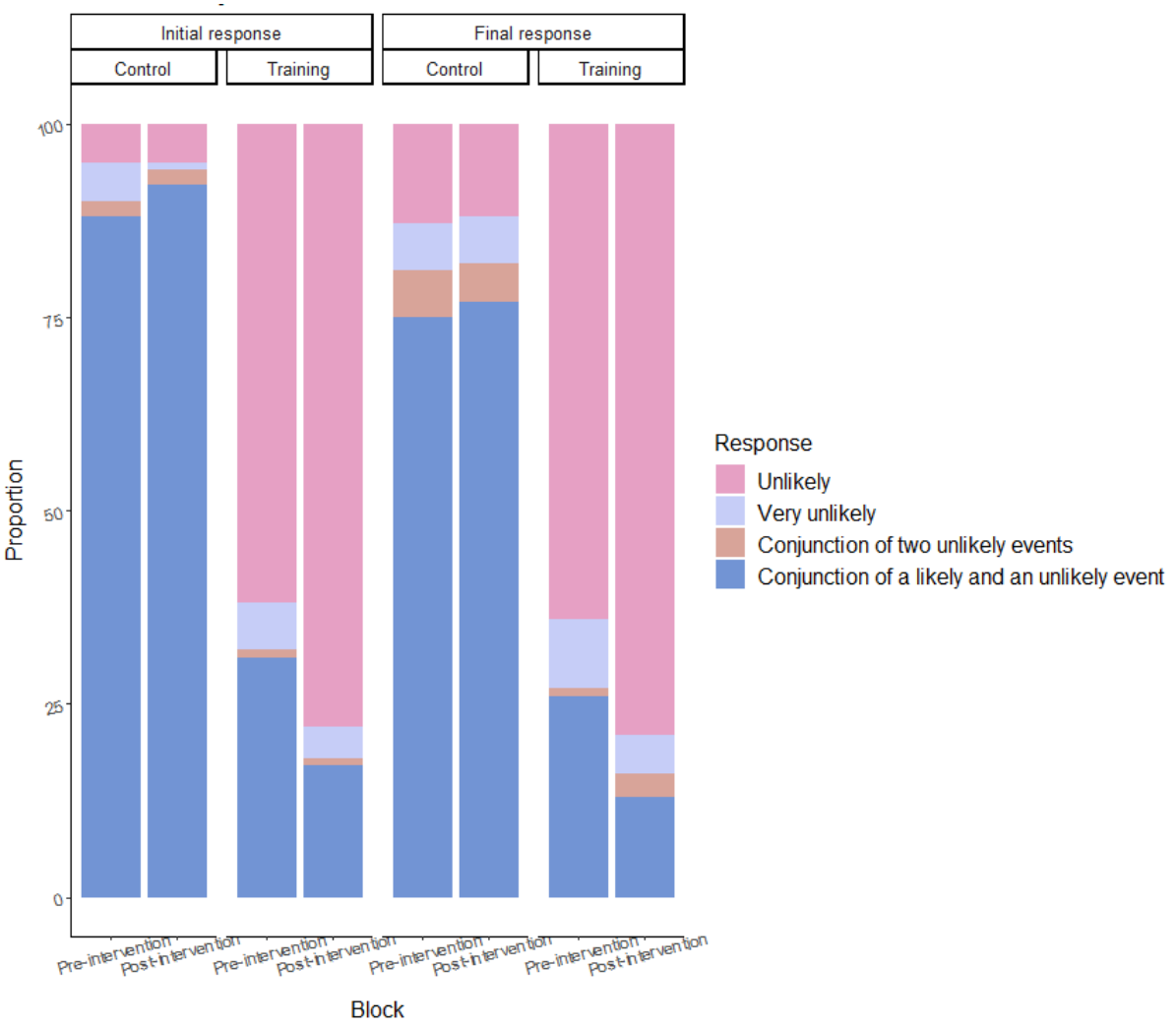


Figure S3. Frequency of each individual response option in Session 2 (conjunction fallacy items) for the initial and the final responses, before and after the intervention in the control and training group.

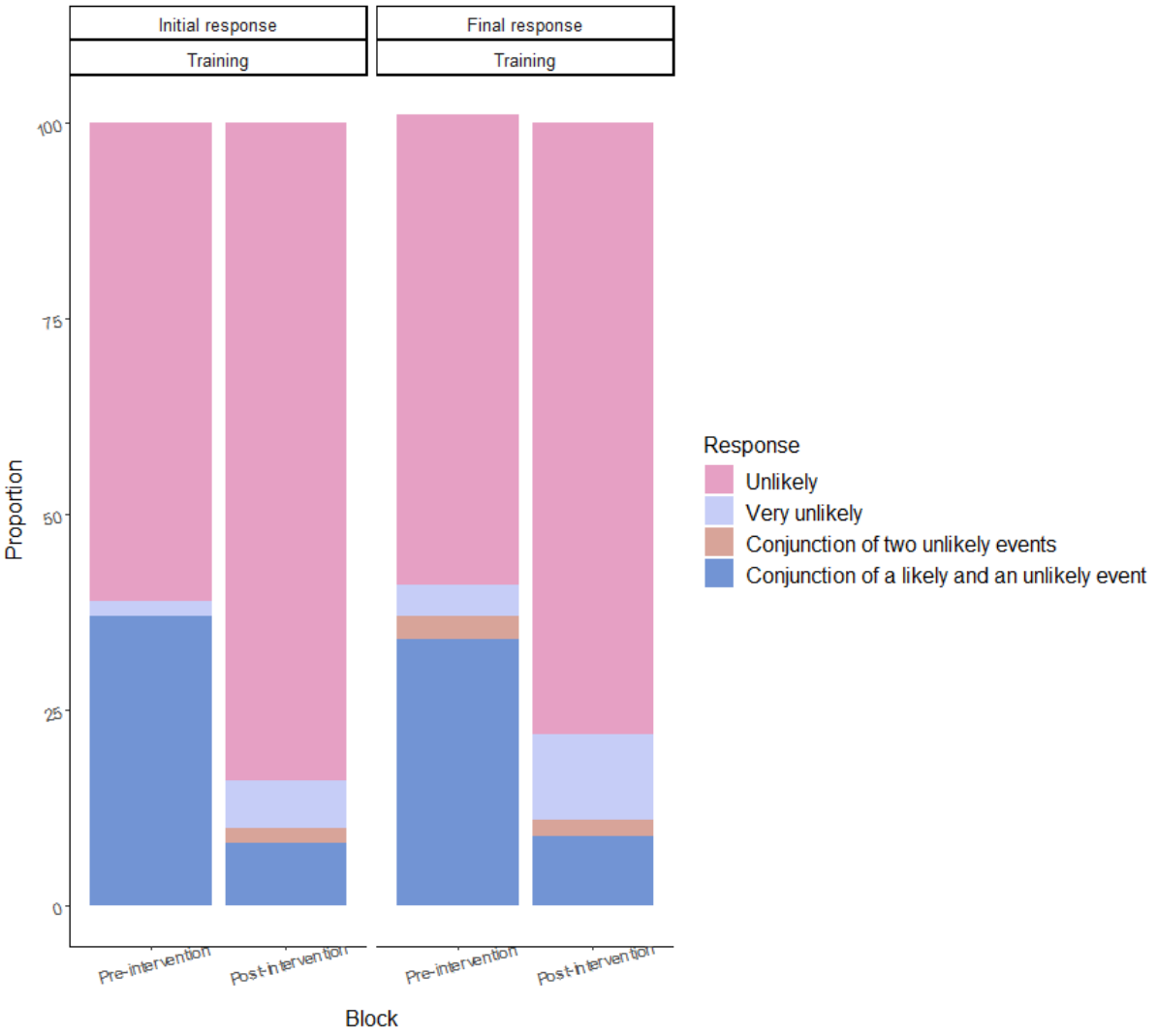


Figure S4. Frequency of each individual response option in Session 3 (conjunction fallacy items) for the initial and the final responses, before and after the intervention in the training group.

D. Accuracy for no-conflict problems in Session 1, Session 2 and Session 3

Table S1.

Average accuracy (%) for the no-conflict problems (SD) of bat-and-ball (BB), base-rate (BR) and conjunction fallacy (CF) tasks and combined (All task) in Session 1.

Task	Group	Initial response - Session 1		Final response - Session 1	
		Pre-intervention	Post-intervention	Pre-intervention	Post-intervention
BB	Control	89.5 (23.5)	93.5 (18.6)	94.2 (17.9)	97.8 (14.7)
	Training	94.0 (14.5)	89.0 (26.2)	97.5 (8.9)	91.7 (24.9)
BR	Control	99.3 (5.0)	93.1 (14.9)	99.3 (5.0)	95.3 (12.9)
	Training	98.4 (6.8)	96.7 (13.6)	96.6 (10.1)	96.8 (11.2)
CF	Control	75.2 (29.2)	64.0 (29.7)	73.6 (29.4)	72.1 (34.3)
	Training	78.6 (22.0)	90.7 (19.4)	81.2 (21.5)	92.9 (19.4)
All task	Control	86.6 (15.0)	82.8 (14.3)	88.6 (12.5)	87.5 (13.7)
	Training	89.6 (10.1)	92.1 (13.0)	91.2 (9.2)	93.4 (13.9)

Table S2.

Average accuracy (%) for the no-conflict problems (SD) of bat-and-ball (BB), base-rate (BR) and conjunction fallacy (CF) tasks and combined (All task) in Session 2.

Task	Group	Initial response - Session 2		Final response - Session 2	
		Pre-intervention	Post-intervention	Pre-intervention	Post-intervention
BB	Control	85.7 (24.2)	96.1 (18.1)	85.9 (22.8)	95.9 (16.3)
	Training	82.3 (28.2)	90.4 (22.7)	81.8 (24.6)	95.0 (19.9)
BR	Control	94.0 (16.0)	92.7 (19.6)	97.4 (8.7)	97.4 (8.4)
	Training	92.8 (19.7)	83.8 (20.8)	96.3 (17.5)	86.4 (17.8)
CF	Control	76.5 (31.3)	65.8 (31.3)	78.0 (31.7)	74.6 (35.5)
	Training	88.7 (23.1)	94.3 (19.0)	93.0 (17.3)	92.1 (21.7)
All task	Control	85.7 (16.5)	84.7 (16.3)	87.4 (11.9)	89.2 (13.9)
	Training	88.2 (15.1)	89.2 (12.0)	90.1 (12.6)	91.1 (11.1)

Table S3.

Average accuracy (%) for the no-conflict problems (SD) of bat-and-ball (BB), base-rate (BR) and conjunction fallacy (CF) tasks and combined (All task) in Session 3.

Task	Group	Initial response - Session 3		Final response - Session 3	
		Pre-intervention	Post-intervention	Pre-intervention	Post-intervention
BB	Training	87.4 (24.7)	89.5 (23.7)	96.6 (12.8)	92.2 (24.7)
BR	Training	98.5 (6.0)	98.8 (5.8)	99.0 (4.9)	100.0 (0)
CF	Training	81.8 (32.5)	93.0 (15.2)	82.0 (32.5)	94.0 (14.8)
All task	Training	89.0 (14.0)	93.2 (11.3)	91.5 (14.7)	95.0 (11.4)

E. Direction of change by task in Session 1, Session 2 and Session 3

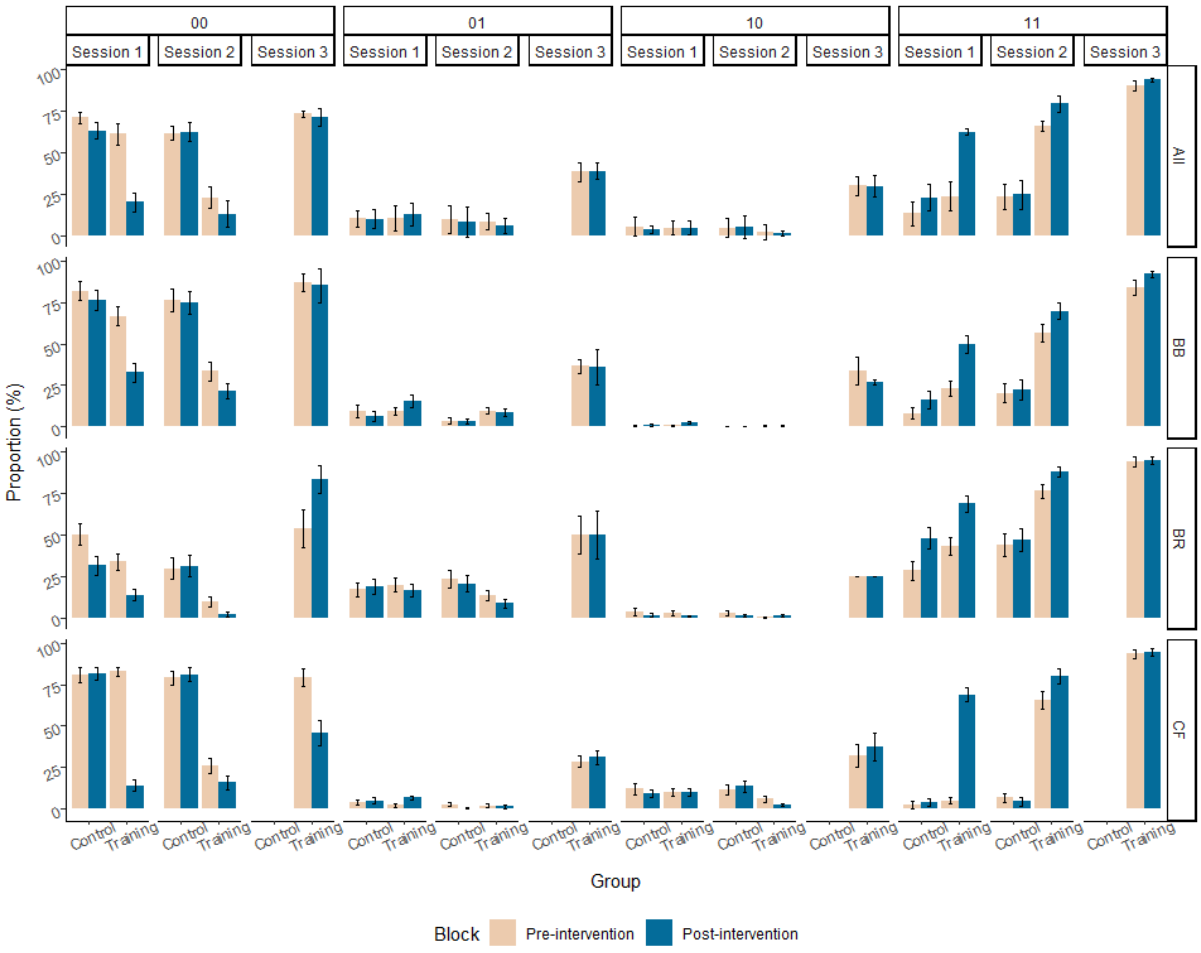
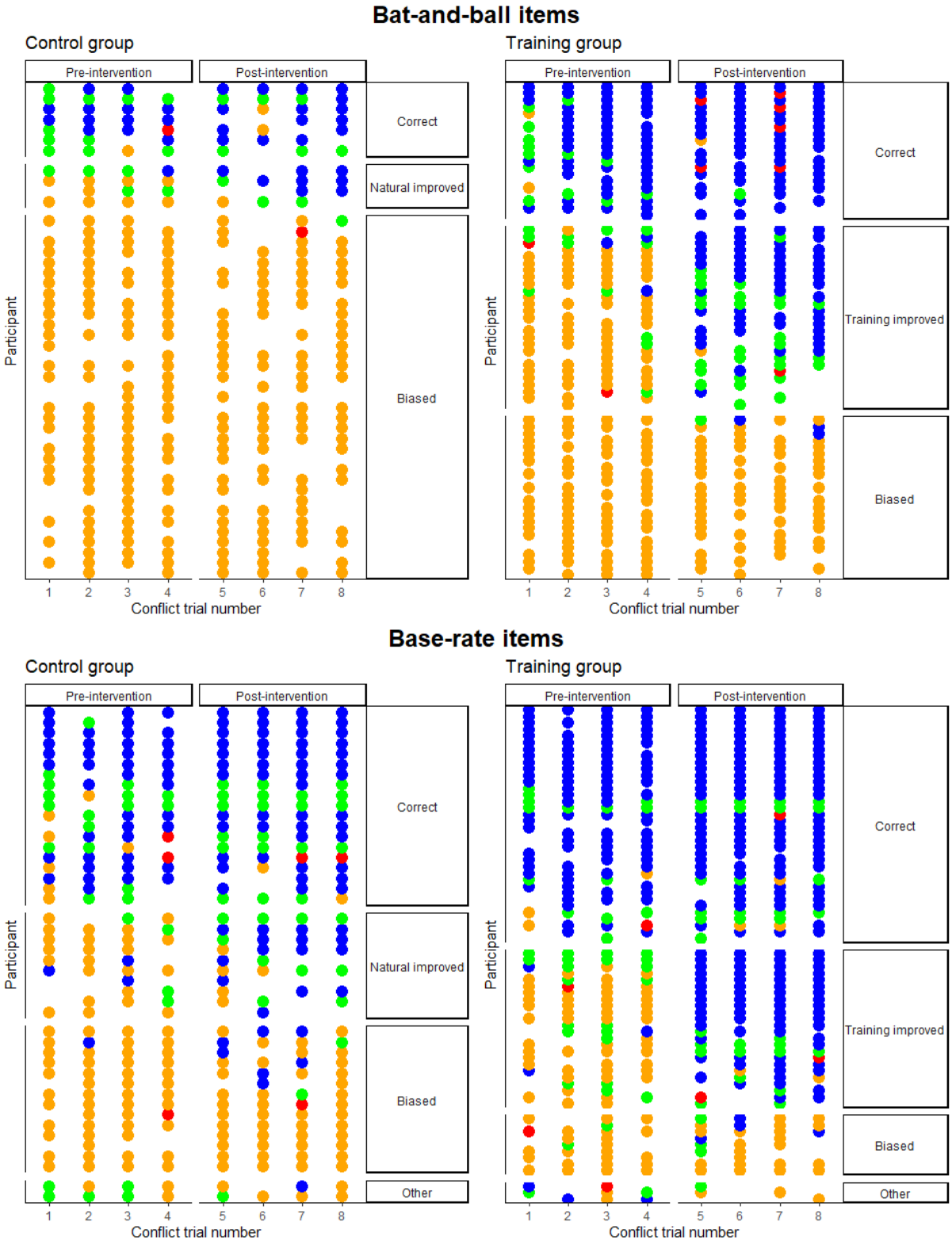


Figure S5. Proportion (%) of each direction of change (i.e., "00" trials, "01" trials, "10" trials and "11" trials) for the conflict problems according to block, group, and task in Session 1, Session 2, and Session 3, for each task (BB, BR, CF), and combined (All). Error bars are standard errors. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks, All = the composite mean across the three tasks.

F. Individual level direction of change in Session 1, Session 2 and Session 3



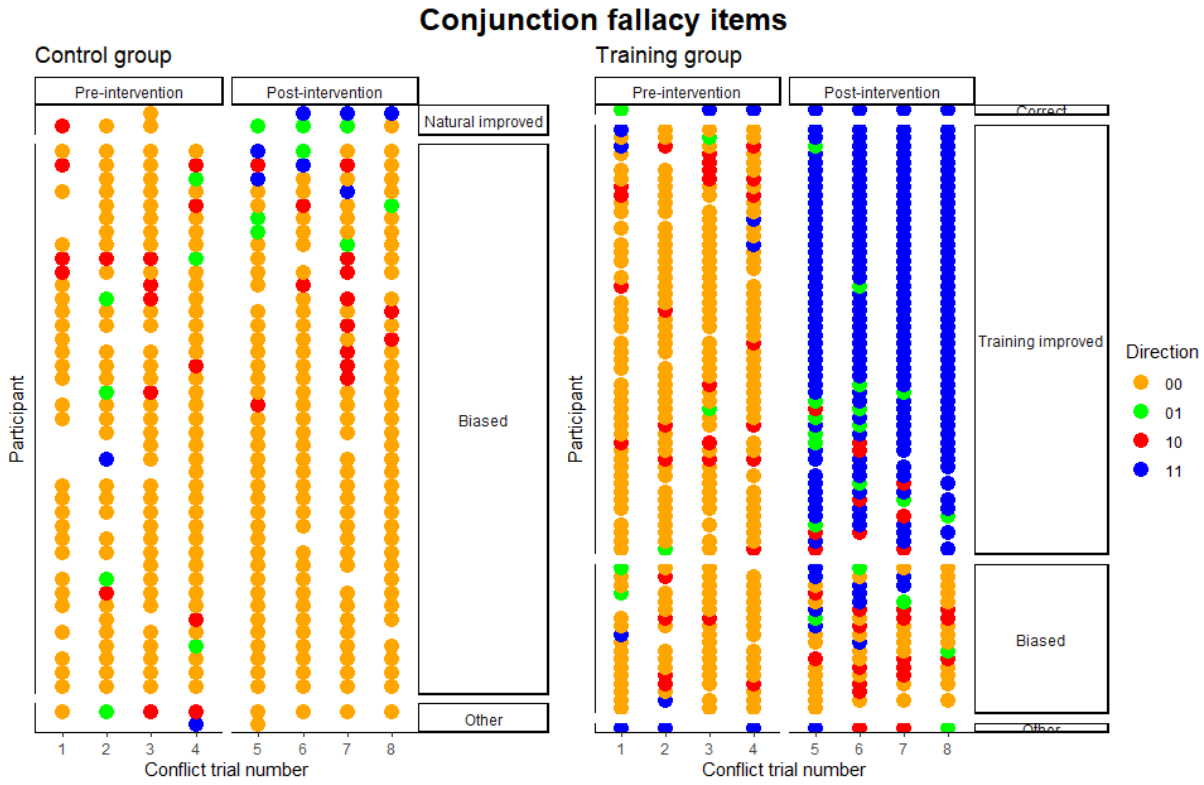
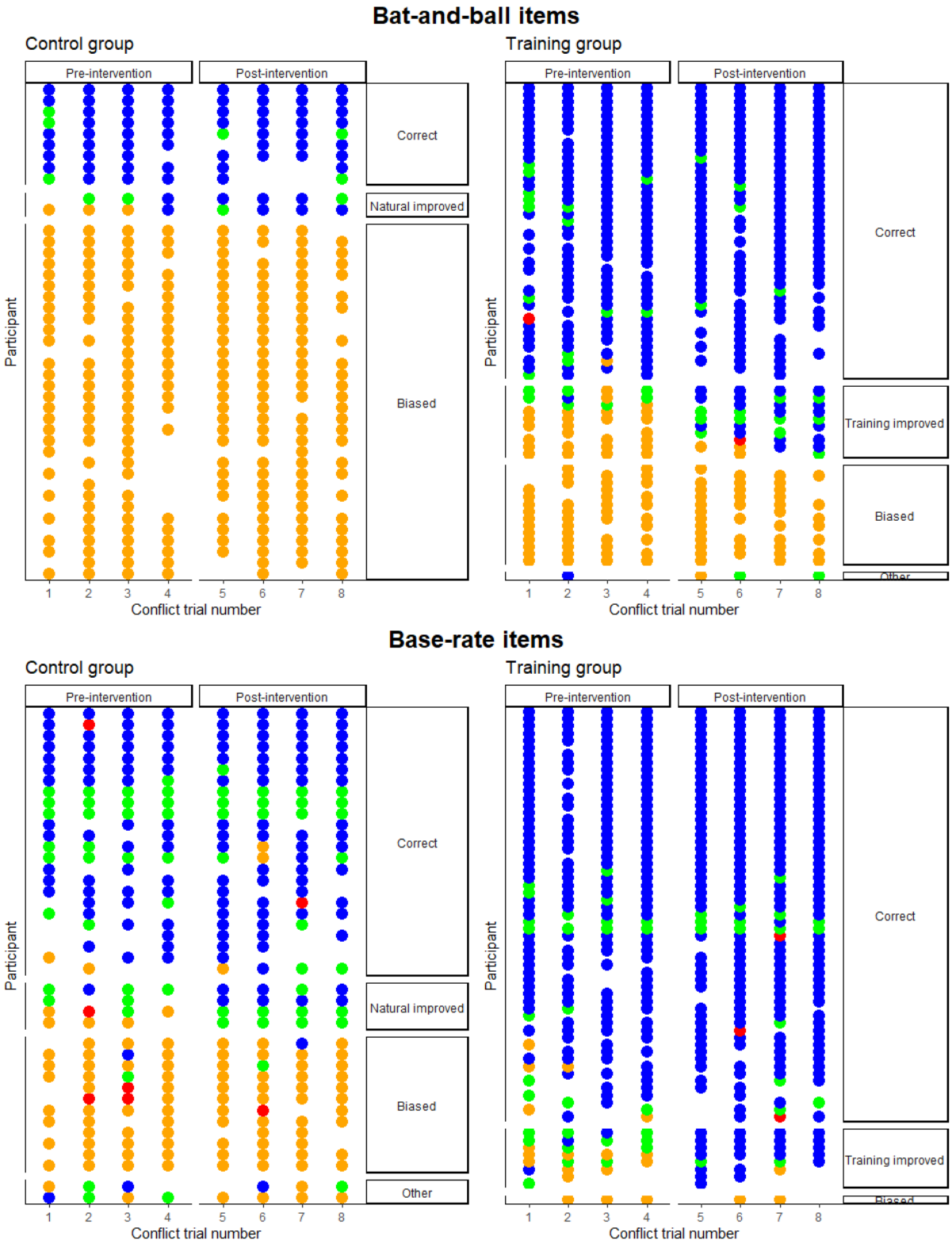


***Figure S6*.** Individual level direction of change (each row represents one participant) and classification in Session 1. Due to the exclusion of missed deadline and load trials (see Trial Exclusion), not all participants contributed 24 analysable trials.


Bat-and-ball items
Control group
Training group
Pre-intervention
Post-intervention
Correct
Natural improved
Biased
Training improved
Other
Participant
Conflict trial number
1
2
3
4
5
6
7
8
Base-rate items
Control group
Training group
Pre-intervention
Post-intervention
Correct
Natural improved
Biased
Training improved
Other
Participant
Conflict trial number

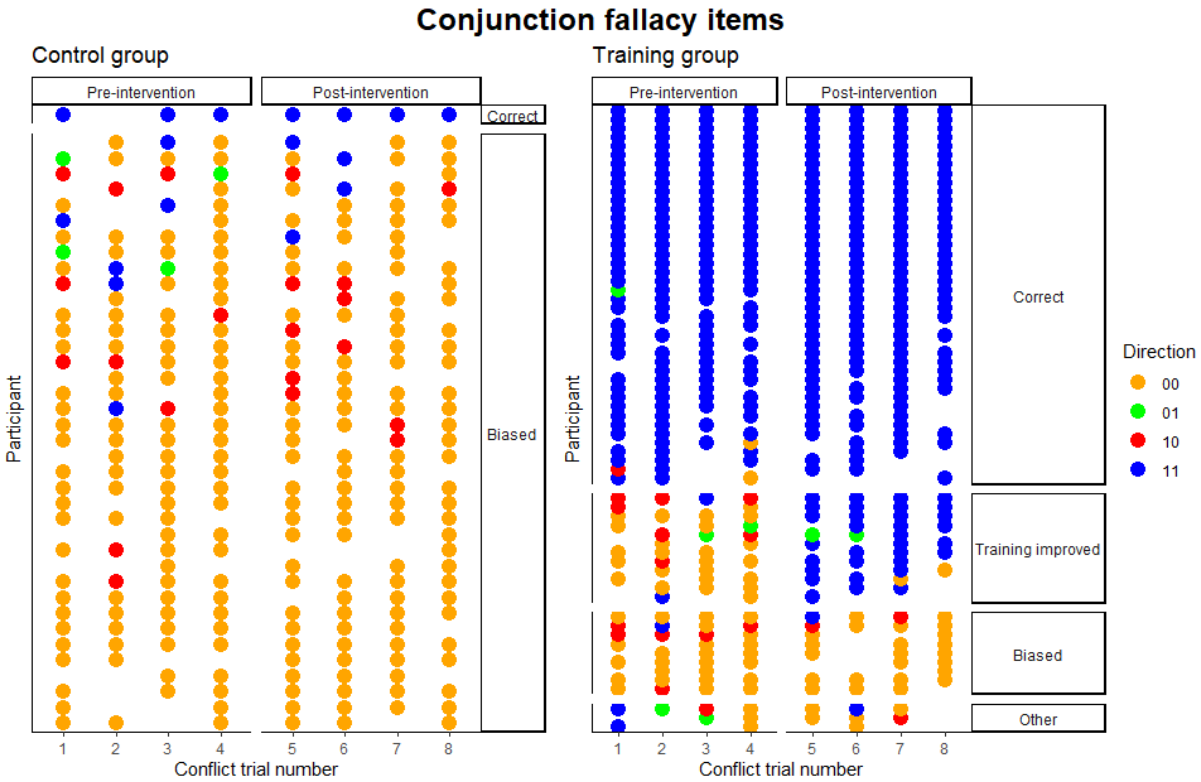
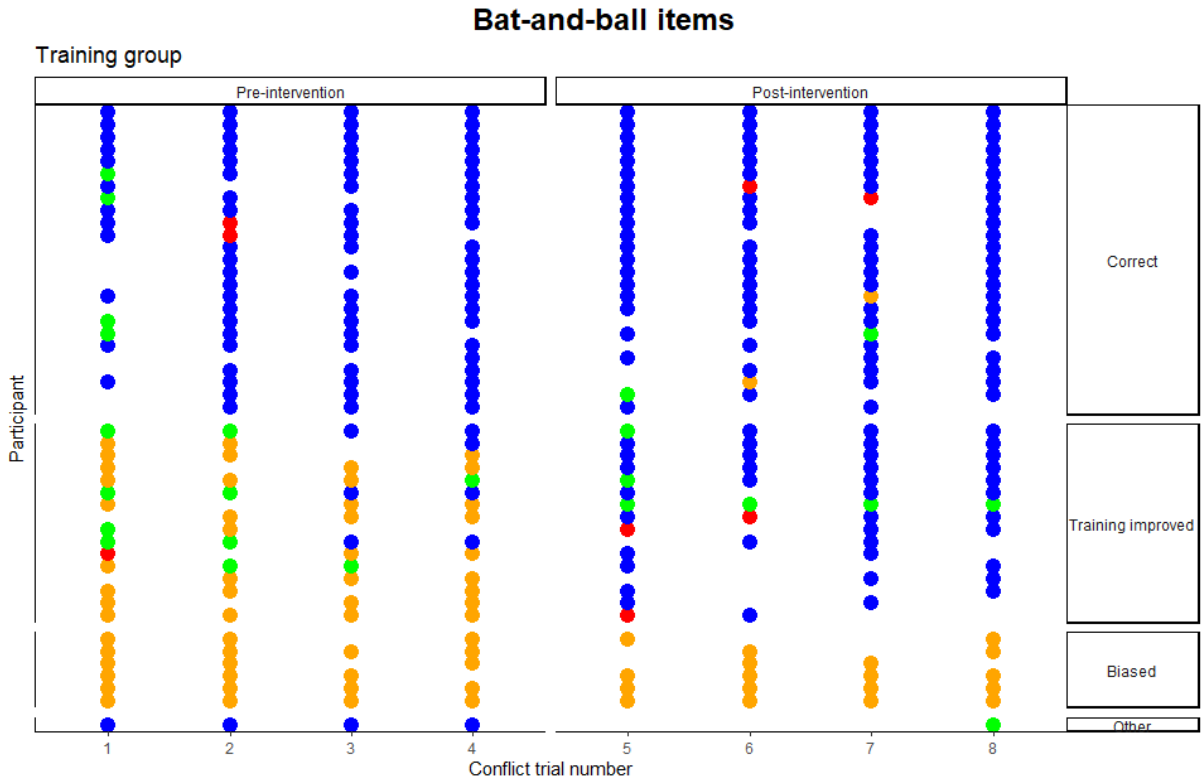
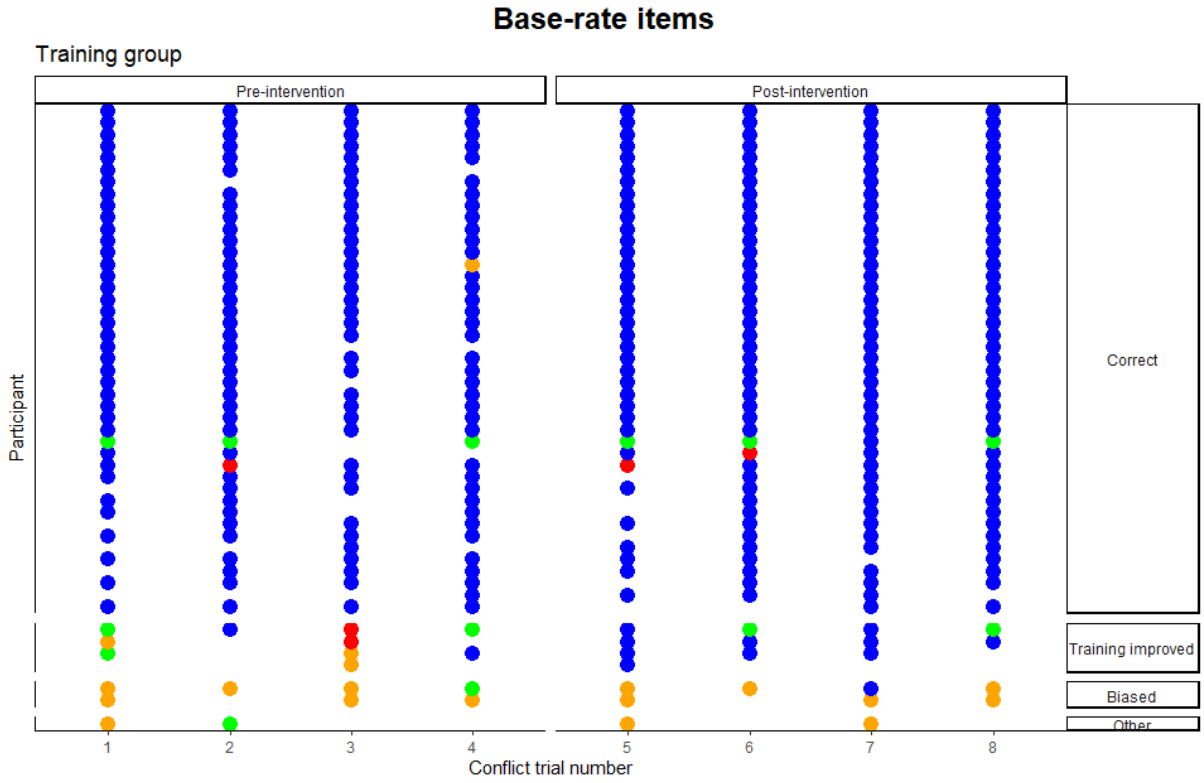


***Figure S7*.** Individual level direction of change (each row represents one participant) and classification in Session 2. Due to the exclusion of missed deadline and load trials (see Trial Exclusion), not all participants contributed 24 analysable trials.


Bat-and-ball items
Training group
Pre-intervention
Post-intervention
Correct
Training improved
Biased
Other
Participant
1
2
3
4
5
6
7
8
Conflict trial number


Base-rate items
Training group
Pre-intervention
Post-intervention
Correct
Training improved
Biased
Other
Participant
1
2
3
4
5
6
7
8
Conflict trial number

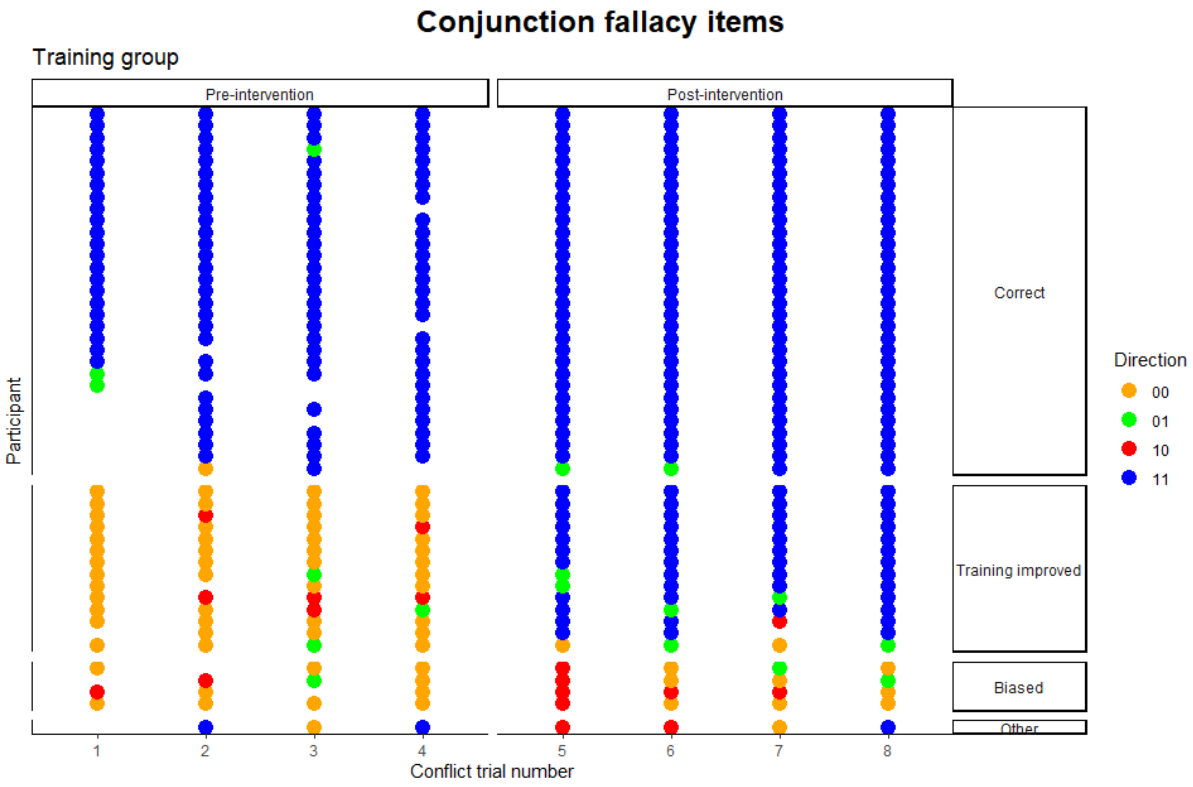


***Figure S8*.** Individual level direction of change (each row represents one participant) and classification in Session 3. Due to the exclusion of missed deadline and load trials (see Trial Exclusion), not all participants contributed 24 analysable trials.

G. Order effect in Session 1, Session 2 and Session 3

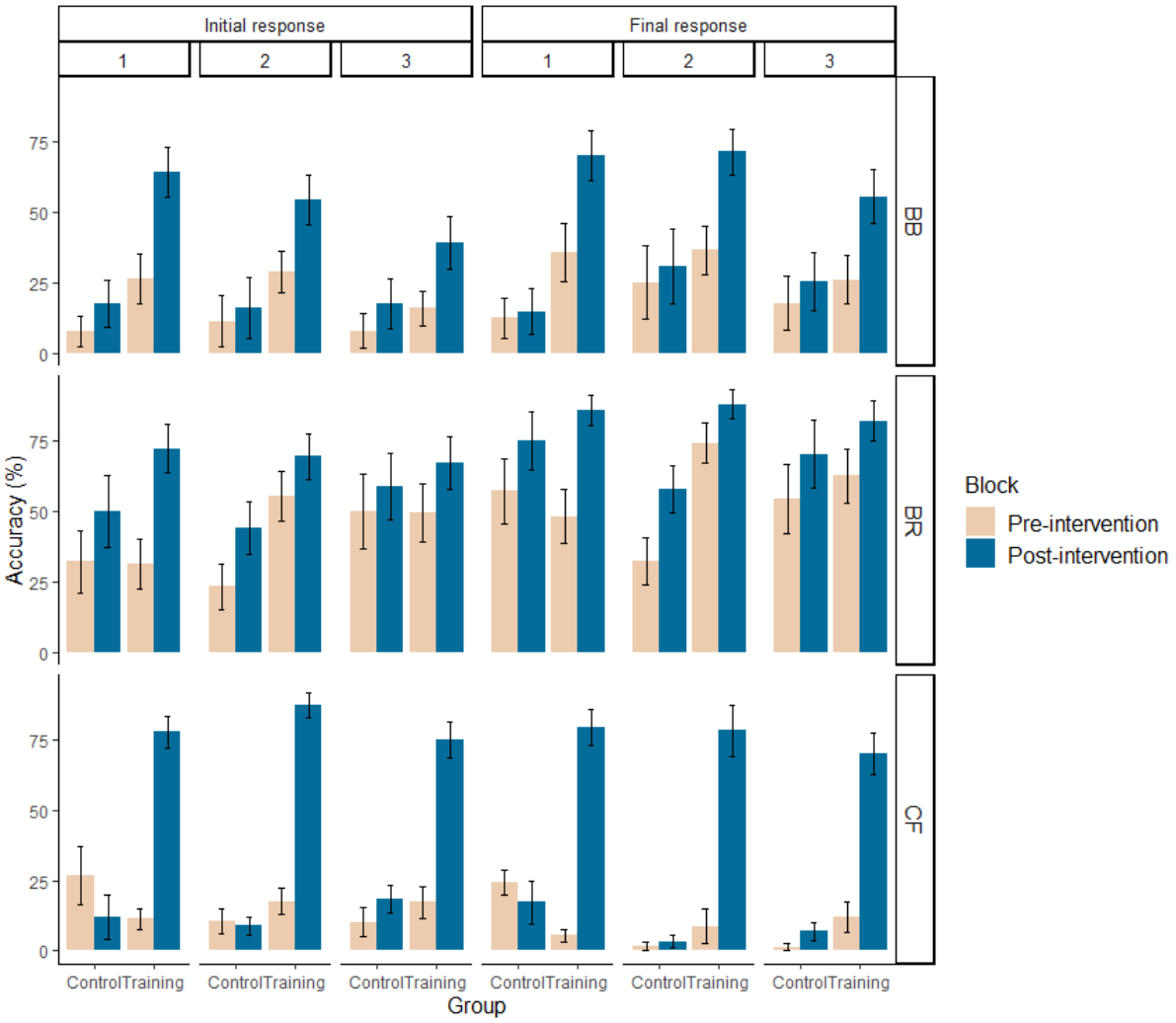


Figure S9. Accuracy on conflict items according to task order in Session 1 in the control and training group. Task order have been randomized. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks.

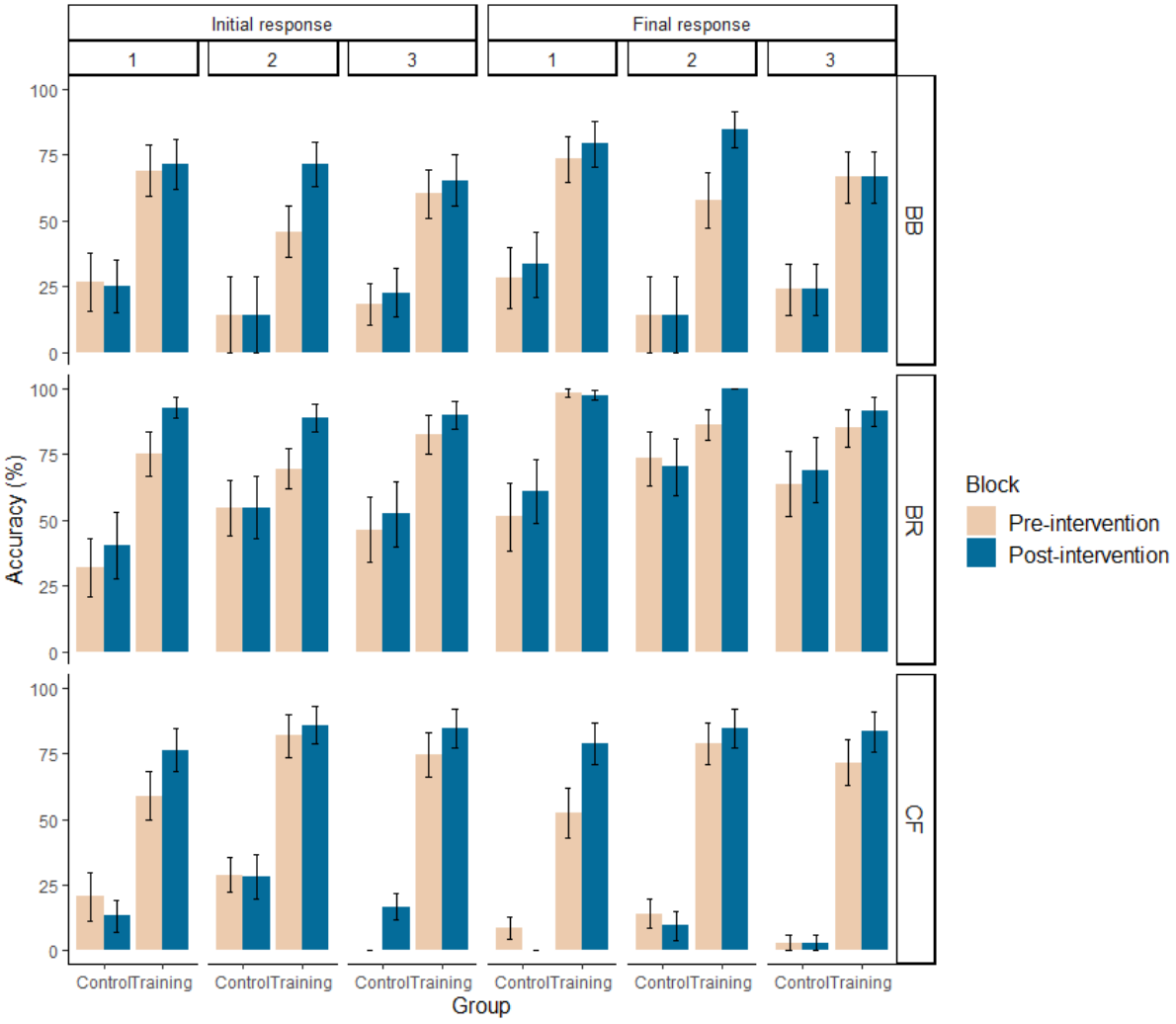


***Figure S10*.** Accuracy on conflict items according to task order in Session 2 in the control and training group. Task order have been randomized. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks.

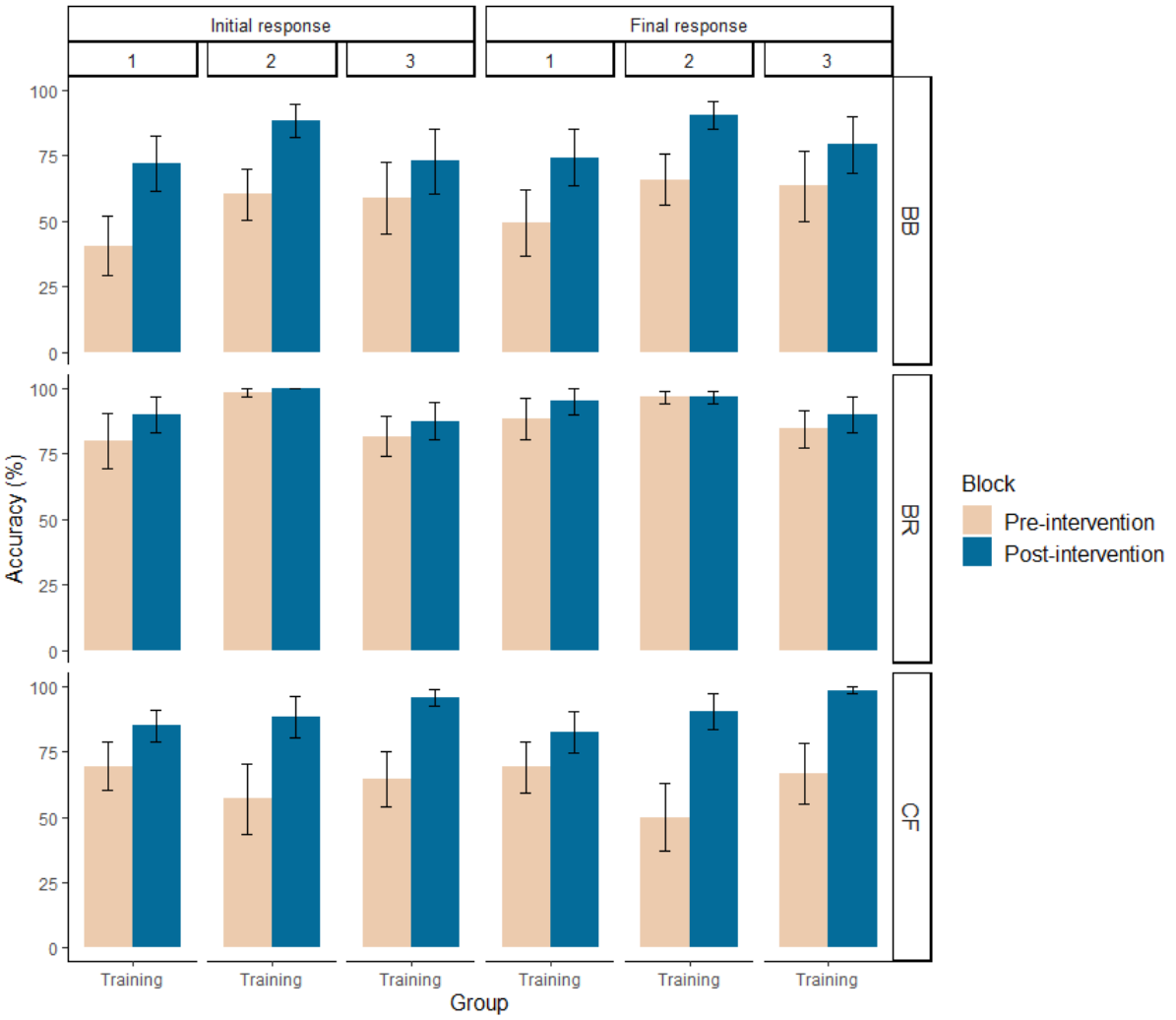


Figure S11. Accuracy on conflict items according to task order in Session 3 in the training group. Task order have been randomized. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy problems.

H. Conflict detection confidence in Session 1, Session 2 and Session 3

Study 1:

Table S4.

Conflict detection results in Session 1 and Session 2. Percentage of mean difference in confidence ratings (SD) between correct no-conflict and incorrect conflict problems on each reasoning task: Bat-and-ball (BB), base-rate neglect (BR) and conjunction fallacy (CF).

Task	Group	Initial response - Session 1		Initial response - Session 2	
		Pre-intervention	Post-intervention	Pre-intervention	Post-intervention
BB	Control	5.7 (24.0)	4.4 (17.8)	-1.1 (23.0)	0.5 (13.8)
	Training	7.7 (20.9)	18.3 (31.7)	4.5 (22.7)	5.6 (19.8)
BR	Control	11.7 (26.1)	9.1 (21.0)	11.8 (20.3)	6.4 (16.1)
	Training	5.8 (18.0)	10.4 (25.3)	16.6 (31.0)	24.0 (34.9)
CF	Control	4.0 (17.2)	3.6 (16.1)	0.9 (20.8)	4.3 (16.2)
	Training	15.2 (18.1)	10.4 (19.8)	-0.6 (32.2)	-1.0 (22.0)

Table S5.

Predictive Conflict Detection results in Session 1 and Session 2. Percentage of mean difference in confidence rating (SD) between correct no-conflict and incorrect conflict problems in the pre-intervention block, for biased vs improved reasoners of the training group, and for each reasoning task: Bat-and-ball (BB), base-rate neglect (BR) and conjunction fallacy (CF).

Task	Label	Initial response – Session 1	Initial response – Session 2
		Pre-intervention	Pre-intervention
BB	Improved	7.9 (18.2)	36.3 (32.8)
	Biased	4.8 (23.3)	0.3 (17.3)
BR	Improved	8.2 (19.0)	31.3 (36.9)
	Biased	4.0 (12.9)	8.2 (11.5)
CF	Improved	7.6 (15.7)	7.6 (21.8)
	Biased	16.2 (16.2)	4.7 (26.6)

Study 2:

Table S6.

Conflict detection results in Session 3 for the training group. Percentage of mean difference in confidence ratings (SD) between correct no-conflict and incorrect conflict problems on each reasoning task: Bat-and-ball (BB), base-rate neglect (BR), conjunction fallacy (CF).

Task	Group	Initial response - Session 3	
		Pre-intervention	Post-intervention
BB	Training	16.1 (28.4)	26.9 (33.8)
BR	Training	4.0 (8.0)	14.6 (33.8)
CF	Training	9.4 (15.9)	21.8 (30.9)

Table S7.

Predictive Conflict Detection results in Session 3. Percentage of mean difference in confidence rating (SD) between correct no-conflict and incorrect conflict problems in the pre-intervention block, for biased vs improved reasoners of the training group, and for each reasoning task: Bat-and-ball (BB), base-rate neglect (BR) and conjunction fallacy (CF).

Task	Label	Initial response
		Pre-intervention
BB	Improved	23.5 (23.8)
	Biased	11.6 (20.4)
BR	Improved	20.0 (26.2)
	Biased	7.0 (9.6)
CF	Improved	29.2 (42.9)
	Biased	5.3 (13.3)

I. Justifications in Session 1, Session 2 and Session 3

Table S8.

Frequency of different types of justifications for the final bat-and-ball (BB), base-rate (BR), conjunction fallacy (CF) conflict problems and all tasks combined (All) during the post-intervention in Session 1.

Task	**Justification – Session 1**	**Control group**		**Training group**	
		Correct response (n = 44)	*Incorrect response (n = 94)*	*Correct response (n=171)*	*Incorrect response (n=50)*
All	Math - Correct	26	-	112	
	Math – Incorrect/Unspecified	3	32	12	17
	Guess	2	12	10	7
	Intuitions	7	30	22	21
	Other	6	20	15	5
BB	Math - Correct	9	-	42	-
	Math – Incorrect/Unspecified	1	22	1	9
	Guess	-	1	2	4
	Intuitions	1	6	3	8
	Other	1	5	3	2
BR	Math - Correct	17	-	54	-
	Math – Incorrect/Unspecified	-	3	2	2
	Guess	1	-	2	-
	Intuitions	6	9	6	6
	Other	5	5	-	2
CF	Math - Correct	-	-	16	-
	Math – Incorrect/Unspecified	2	7	9	6
	Guess	1	11	6	3
	Intuitions	-	15	13	7
	Other	-	10	12	1

Table S9.

Frequency of different types of justifications for the final bat-and-ball (BB), base-rate (BR), conjunction fallacy (CF) conflict problems and all tasks combined (All) during the post-intervention in Session 2.

Task	**Justification – Session 2**	**Control group**		**Training group**	
		Correct response (n = 13)	*Incorrect response (n = 73)*	*Correct response (n = 109)*	*Incorrect response (n = 25)*
All	Math - Correct	12	-	66	-
	Math – Incorrect/Unspecified	-	33	6	10
	Guess	1	10	5	3
	Intuitions	-	20	18	12
	Other	-	10	14	-
BB	Math - Correct	11	-	44	-
	Math – Incorrect/Unspecified	-	22	1	7
	Guess	-	2	3	1
	Intuitions	-	6	6	4
	Other	-	2	2	-
CF	Math - Correct	1	-	22	-
	Math – Incorrect/Unspecified	-	11	5	3
	Guess	1	8	5	2
	Intuitions	-	14	12	8
	Other	-	8	12	-

Note. Due to a coding error, justification data for the base-rate task is missing in Session 2 (see Justification in the Material section).

Table S10.

Frequency of different types of justifications for the final bat-and-ball (BB), base-rate (BR), conjunction fallacy (CF) conflict problems and all tasks combined (All) during the post-intervention in Session 3.

Task	**Justification – Session 3**	**Training group**	
		Correct response (n = 137)	*Incorrect response (n = 13)*
All	Math - Correct	92	-
	Math – Incorrect/Unspecified	7	5
	Guess	5	3
	Intuitions	19	4
	Other	14	1
BB	Math - Correct	32	-
	Math – Incorrect/Unspecified	3	4
	Guess	1	2
	Intuitions	4	2
	Other	2	-
BR	Math - Correct	36	-
	Math – Incorrect/Unspecified	-	-
	Guess	1	-
	Intuitions	7	1
	Other	4	1
CF	Math - Correct	24	-
	Math – Incorrect/Unspecified	4	1
	Guess	3	1
	Intuitions	8	1
	Other	8	-

J. Comparison between delayed training interventions

Table S11.

Comparison between post-intervention mean accuracies (%) for the conflict problems (SD) and decay in performance (%) after two months, for each task (BB, BR, CF), and combined (All), in single training session in Boissin et al.'s (2021, 2022) studies, and after repeated training sessions in the current study. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy, All = the composite mean across the three tasks.

Task	Post-intervention mean accuracies (%) for the conflict problems (SD) obtained after two months				Decay (%) between last post-test and +two months pre-test			
	Boissin et al. study		Current study		Boissin et al. study		Current study	
	Initial	Final	Initial	Final	Initial	Final	Initial	Final
All	68.6 (39.8)	73.9 (38.0)	87.8 (21.3)	89.2 (21.5)	19	17	12	14
BB	69.5 (6.5)	75.0 (6.1)	79.7 (10.8)	83.0 (9.8)	18	15	15	16
BR	75.6 (6.6)	84.9 (6.0)	92.0 (7.1)	93.5 (6.4)	18	6	4	7
CF	66.2 (7.1)	65.8 (7.1)	89.5 (6.7)	90.0 (7.6)	20	29	17	19

Note. For the post-intervention mean accuracies obtained after two months, it corresponds to Study 3 for bat-and-ball in Boissin et al. (2021), Studies 3 and 4 in Boissin et al. (2022) for respectively base-rate and conjunction fallacy. In the current study, it corresponds to Study 2 post-intervention block. For the decay between two months, i.e., the difference between the delayed pre-intervention and last post-intervention blocks, this implies post-intervention Session 1 and pre-intervention Session 2 (re-test) in Boissin et al. (2021, 2022) studies and post-intervention Session 2 and pre-intervention Session 3 in the current study.

K. Supplementary statistics

Correlation of all variables of each condition in Session 1:

Table S12.

Correlation table in the training group in Session 1 showing Pearson correlation coefficients (r) between task (BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks), blocks (pre- and post-intervention) and response stages (initial and final). Sample size was 74.

Variables	BB_Post _Final	BB_Post _Initial	BB_Pre _Final	BB_Pre _Initial	BR_Post _Final	BR_Post _Initial	BR_Pre _Final	BR_Pre _Initial	CF_Post _Final	CF_Post _Initial	CF_Pre _Final	CF_Pre _Initial
BB_Post_Final	**1.00**	0.72	0.48	0.42	0.34	0.38	0.28	0.24	0.28	0.37	-0.23	-0.27
BB_Post_Initial	0.72	**1.00**	0.74	0.68	0.31	0.51	0.31	0.45	0.5	0.49	-0.11	-0.22
BB_Pre_Final	0.48	0.74	**1.00**	0.92	0.27	0.39	0.33	0.39	0.36	0.38	0.1	-0.12
BB_Pre_Initial	0.42	0.68	0.92	**1.00**	0.25	0.39	0.3	0.38	0.36	0.36	0.05	-0.09
BR_Post_Final	0.34	0.31	0.27	0.25	**1.00**	0.60	0.39	0.32	0.48	0.61	-0.09	-0.05
BR_Post_Initial	0.38	0.51	0.39	0.39	0.60	**1.00**	0.32	0.49	0.40	0.64	-0.06	-0.08
BR_Pre_Final	0.28	0.31	0.33	0.3	0.39	0.32	**1.00**	0.65	0.48	0.48	0.02	-0.12
BR_Pre_Initial	0.24	0.45	0.39	0.38	0.32	0.49	0.65	**1.00**	0.43	0.46	0.01	-0.04
CF_Post_Final	0.28	0.5	0.36	0.36	0.48	0.40	0.48	0.43	**1.00**	0.79	-0.06	0.03
CF_Post_Initial	0.37	0.49	0.38	0.36	0.61	0.64	0.48	0.46	0.79	**1.00**	-0.04	-0.03
CF_Pre_Final	-0.23	-0.11	0.1	0.05	-0.09	-0.06	0.02	0.01	-0.06	-0.04	**1.00**	0.37
CF_Pre_Initial	-0.27	-0.22	-0.12	-0.09	-0.05	-0.08	-0.12	-0.04	0.03	-0.03	0.37	**1.00**

Note. The name of each variable is noted in the general form of "Task_Block_Response stage", describing the task, the block, and the response stage we are looking at. For example, "BB_Post_Final" means we are focusing on final responses of the bat-and-ball task, in the post-intervention block.

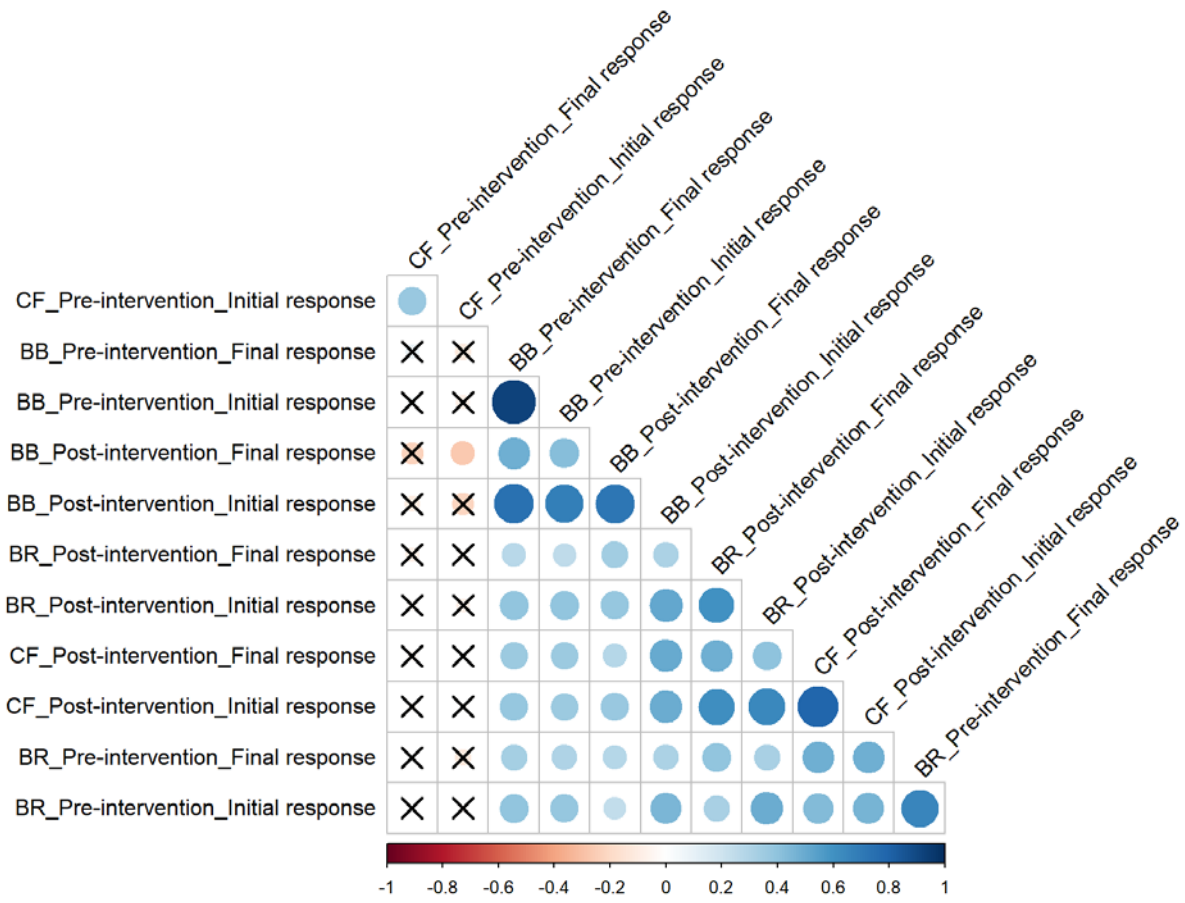


Figure S12. Correlogram in the training group in Session 1: For all task (BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks), blocks (pre- and post-intervention) and response stages (initial and final). Insignificant coefficients are marked with a cross.

Table S13.

Correlation table in the control group in Session 1 showing Pearson correlation coefficients (r) between task (BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks), blocks (pre- and post-intervention) and response stages (initial and final). Sample size was 46.

Variables	BB_Post _Final	BB_Post _Initial	BB_Pre _Final	BB_Pre _Initial	BR_Post _Final	BR_Post _Initial	BR_Pre _Final	BR_Pre _Initial	CF_Post _Final	CF_Post _Initial	CF_Pre _Final	CF_Pre _Initial
BB_Post_Final	**1.00**	0.83	0.87	0.57	0.14	0.12	0.26	0.10	0.03	-0.04	-0.18	-0.17
BB_Post_Initial	0.83	**1.00**	0.80	0.71	-0.03	0.12	0.12	0.12	-0.04	0.03	-0.17	-0.07
BB_Pre_Final	0.87	0.80	**1.00**	0.76	0.11	0.14	0.33	0.18	-0.02	-0.13	-0.16	-0.16
BB_Pre_Initial	0.57	0.71	0.76	**1.00**	0.05	0.12	0.22	0.18	-0.05	-0.12	-0.04	-0.18
BR_Post_Final	0.14	-0.03	0.11	0.05	**1.00**	0.64	0.66	0.52	0.09	0.03	0.04	0.04
BR_Post_Initial	0.12	0.12	0.14	0.12	0.64	**1.00**	0.46	0.72	0.04	0.16	0.15	0.11
BR_Pre_Final	0.26	0.12	0.33	0.22	0.66	0.46	**1.00**	0.67	0.25	0.08	-0.12	-0.20
BR_Pre_Initial	0.10	0.12	0.18	0.18	0.52	0.72	0.67	**1.00**	0.16	0.30	0.17	-0.05
CF_Post_Final	0.03	-0.04	-0.02	-0.05	0.09	0.04	0.25	0.16	**1.00**	0.45	0.05	0.05
CF_Post_Initial	-0.04	0.03	-0.13	-0.12	0.03	0.16	0.08	0.30	0.45	**1.00**	0.05	0.19
CF_Pre_Final	-0.18	-0.17	-0.16	-0.04	0.04	0.15	-0.12	0.17	0.05	0.05	**1.00**	0.31
CF_Pre_Initial	-0.17	-0.07	-0.16	-0.18	0.04	0.11	-0.20	-0.05	0.05	0.19	0.31	**1.00**

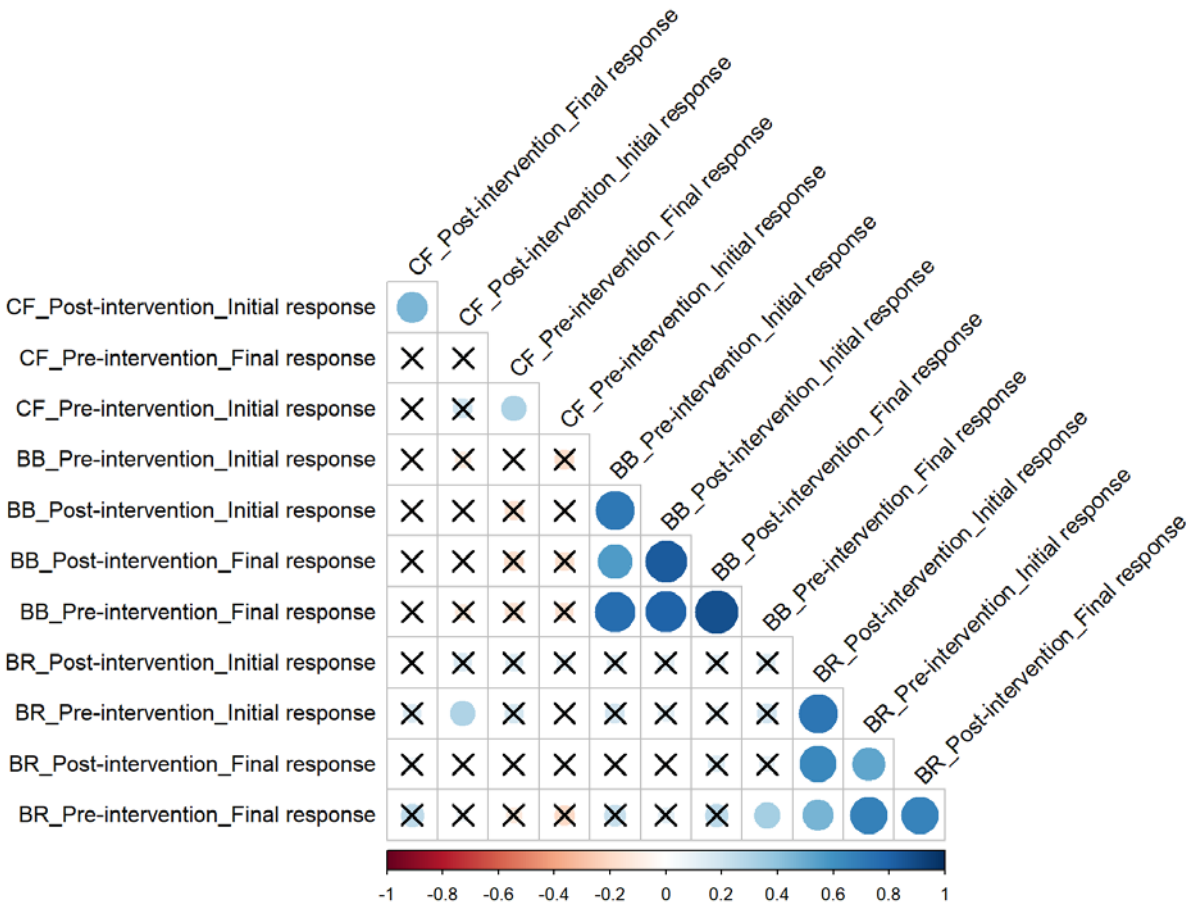


Figure S13. Correlogram in the control group in Session 1: For all task (BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks), blocks (pre- and post-intervention) and response stages (initial and final). Insignificant coefficients are marked with a cross.

Correlation of all variables of each condition in Session 2:

Table S14.

Correlation table in the training group in Session 2 showing Pearson correlation coefficients (r) between task (BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks), blocks (pre- and post-intervention) and response stages (initial and final). Sample size was 67.

Variables	BB_Post _Final	BB_Post _Initial	BB_Pre _Final	BB_Pre _Initial	BR_Post _Final	BR_Post _Initial	BR_Pre _Final	BR_Pre _Initial	CF_Post _Final	CF_Post _Initial	CF_Pre _Final	CF_Pre _Initial
BB_Post_Final	**1.00**	0.88	0.81	0.71	0.20	0.16	0.31	0.31	0.38	0.32	0.34	0.43
BB_Post_Initial	0.88	**1.00**	0.86	0.74	0.32	0.18	0.34	0.35	0.44	0.40	0.40	0.45
BB_Pre_Final	0.81	0.86	**1.00**	0.90	0.31	0.24	0.40	0.45	0.41	0.40	0.48	0.51
BB_Pre_Initial	0.71	0.74	0.90	**1.00**	0.27	0.26	0.37	0.40	0.40	0.39	0.46	0.43
BR_Post_Final	0.20	0.32	0.31	0.27	**1.00**	0.43	0.49	0.26	0.30	0.31	0.26	0.32
BR_Post_Initial	0.16	0.18	0.24	0.26	0.43	**1.00**	0.33	0.56	0.13	0.13	0.03	0.06
BR_Pre_Final	0.31	0.34	0.40	0.37	0.49	0.33	**1.00**	0.68	0.18	0.19	0.39	0.50
BR_Pre_Initial	0.31	0.35	0.45	0.40	0.26	0.56	0.68	**1.00**	0.29	0.27	0.32	0.35
CF_Post_Final	0.38	0.44	0.41	0.40	0.30	0.13	0.18	0.29	**1.00**	0.96	0.63	0.59
CF_Post_Initial	0.32	0.40	0.40	0.39	0.31	0.13	0.19	0.27	0.96	**1.00**	0.65	0.59
CF_Pre_Final	0.34	0.40	0.48	0.46	0.26	0.03	0.39	0.32	0.63	0.65	**1.00**	0.89
CF_Pre_Initial	0.43	0.45	0.51	0.43	0.32	0.06	0.50	0.35	0.59	0.59	0.89	**1.00**

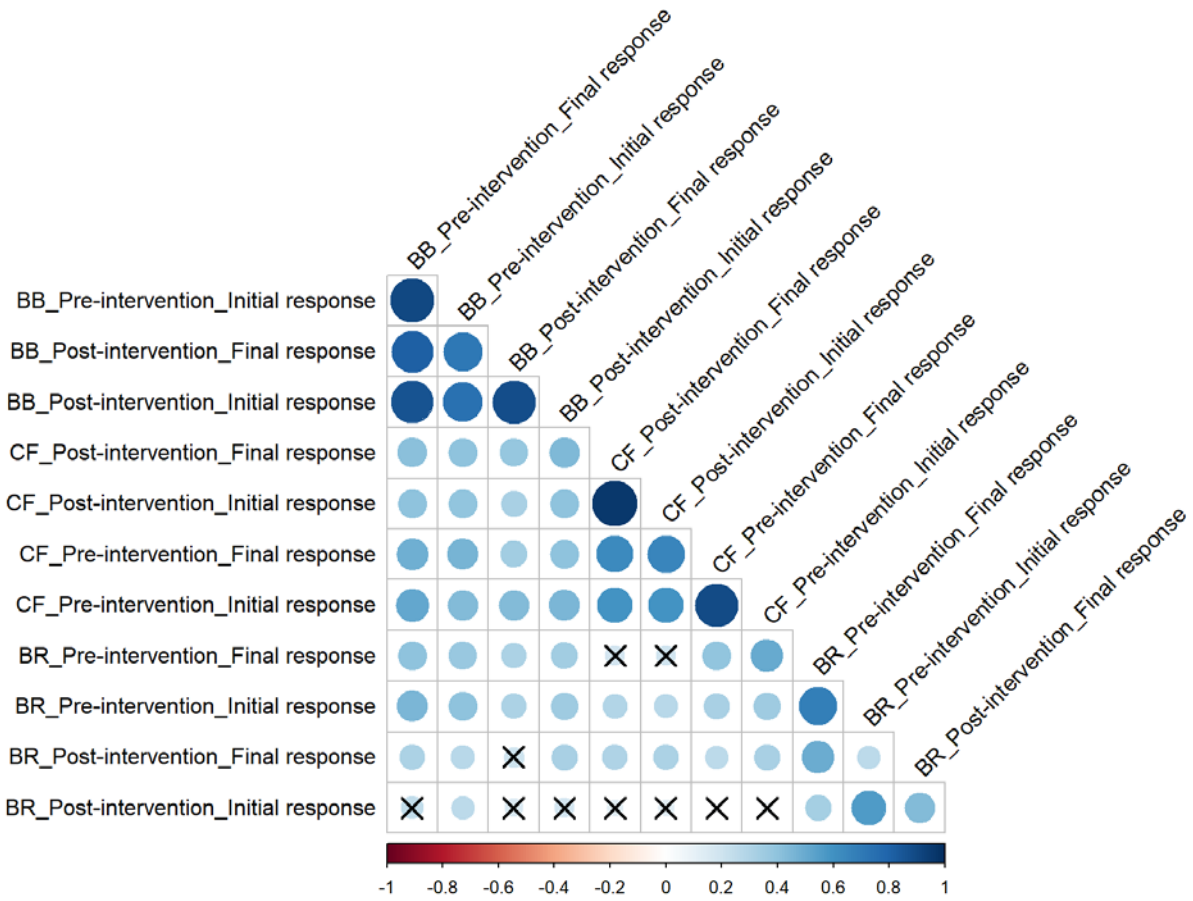


Figure S14. Correlogram in the training group in Session 2: For all task (BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks), blocks (pre- and post-intervention) and response stages (initial and final). Insignificant coefficients are marked with a cross.

Table S15.

Correlation table in the control group in Session 2 showing Pearson correlation coefficients (r) between task (BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks), blocks (pre- and post-intervention) and response stages (initial and final). Sample size was 43.

Variables	BB_Post _Final	BB_Post _Initial	BB_Pre _Final	BB_Pre _Initial	BR_Post _Final	BR_Post _Initial	BR_Pre _Final	BR_Pre _Initial	CF_Post _Final	CF_Post _Initial	CF_Pre _Final	CF_Pre _Initial
BB_Post_Final	**1.00**	0.97	0.97	0.93	0.29	0.18	0.33	0.33	0.14	-0.06	0.05	-0.17
BB_Post_Initial	0.97	**1.00**	0.95	0.93	0.26	0.12	0.30	0.27	0.004	-0.11	-0.05	-0.23
BB_Pre_Final	0.97	0.95	**1.00**	0.96	0.30	0.20	0.33	0.35	0.15	-0.04	0.07	-0.15
BB_Pre_Initial	0.93	0.93	0.96	**1.00**	0.26	0.16	0.30	0.33	0.20	0.002	0.13	-0.11
BR_Post_Final	0.29	0.26	0.30	0.26	**1.00**	0.7	0.80	0.56	0.15	0.06	0.03	-0.16
BR_Post_Initial	0.18	0.12	0.20	0.16	0.7	**1.00**	0.74	0.83	0.09	0.06	0.01	-0.07
BR_Pre_Final	0.33	0.30	0.33	0.30	0.80	0.74	**1.00**	0.62	0.09	0.04	0.02	-0.18
BR_Pre_Initial	0.33	0.27	0.35	0.33	0.56	0.83	0.62	**1.00**	0.11	0.01	0.10	-0.05
CF_Post_Final	0.14	0.004	0.15	0.20	0.15	0.09	0.09	0.11	**1.00**	0.44	0.63	0.36
CF_Post_Initial	-0.06	-0.11	-0.04	0.002	0.06	0.06	0.04	0.01	0.44	**1.00**	0.32	0.62
CF_Pre_Final	0.05	-0.05	0.07	0.13	0.03	0.01	0.02	0.10	0.63	0.32	**1.00**	0.50
CF_Pre_Initial	-0.17	-0.23	-0.15	-0.11	-0.16	-0.07	-0.18	-0.05	0.36	0.62	0.50	**1.00**

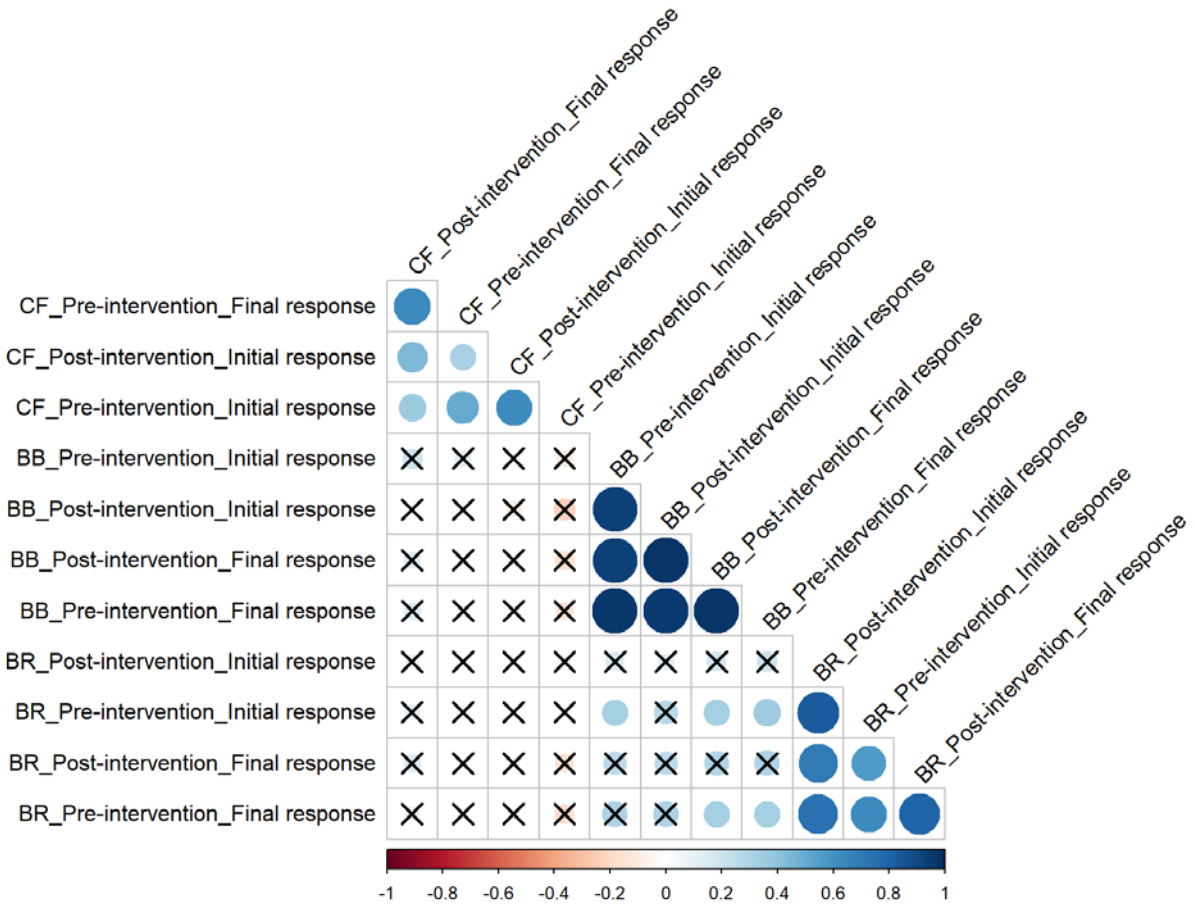


Figure S15. Correlogram in the control group in Session 2: For all task (BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks), blocks (pre- and post-intervention) and response stages (initial and final). Insignificant coefficients are marked with a cross.

Correlation of all variables of each condition in Session 3:

Table S16.

Correlation table in the training group in Session 3 showing Pearson correlation coefficients (r) between task (BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks), blocks (pre- and post-intervention) and response stages (initial and final). Sample size was 50.

Variables	BB_Post _Final	BB_Post _Initial	BB_Pre _Final	BB_Pre _Initial	BR_Post _Final	BR_Post _Initial	BR_Pre _Final	BR_Pre _Initial	CF_Post _Final	CF_Post _Initial	CF_Pre _Final	CF_Pre _Initial
BB_Post_Final	**1.00**	0.76	0.53	0.50	0.56	0.51	0.38	0.38	0.47	0.53	0.31	0.37
BB_Post_Initial	0.76	**1.00**	0.44	0.40	0.48	0.43	0.48	0.34	0.53	0.47	0.46	0.38
BB_Pre_Final	0.53	0.44	**1.00**	0.90	0.29	0.24	0.48	0.38	0.45	0.41	0.51	0.53
BB_Pre_Initial	0.50	0.40	0.90	**1.00**	0.26	0.20	0.45	0.34	0.44	0.40	0.49	0.51
BR_Post_Final	0.56	0.48	0.29	0.26	**1.00**	0.76	0.61	0.57	0.08	0.25	0.18	0.25
BR_Post_Initial	0.51	0.43	0.24	0.20	0.76	**1.00**	0.51	0.78	0.24	0.33	0.20	0.26
BR_Pre_Final	0.38	0.48	0.48	0.45	0.61	0.51	**1.00**	0.81	0.45	0.35	0.24	0.23
BR_Pre_Initial	0.38	0.34	0.38	0.34	0.57	0.78	0.81	**1.00**	0.41	0.44	0.18	0.20
CF_Post_Final	0.47	0.53	0.45	0.44	0.08	0.24	0.45	0.41	**1.00**	0.76	0.41	0.41
CF_Post_Initial	0.53	0.47	0.41	0.40	0.25	0.33	0.35	0.44	0.76	**1.00**	0.45	0.48
CF_Pre_Final	0.31	0.46	0.51	0.49	0.18	0.20	0.24	0.18	0.41	0.45	**1.00**	0.93
CF_Pre_Initial	0.37	0.38	0.53	0.51	0.25	0.26	0.23	0.20	0.41	0.48	0.93	**1.00**

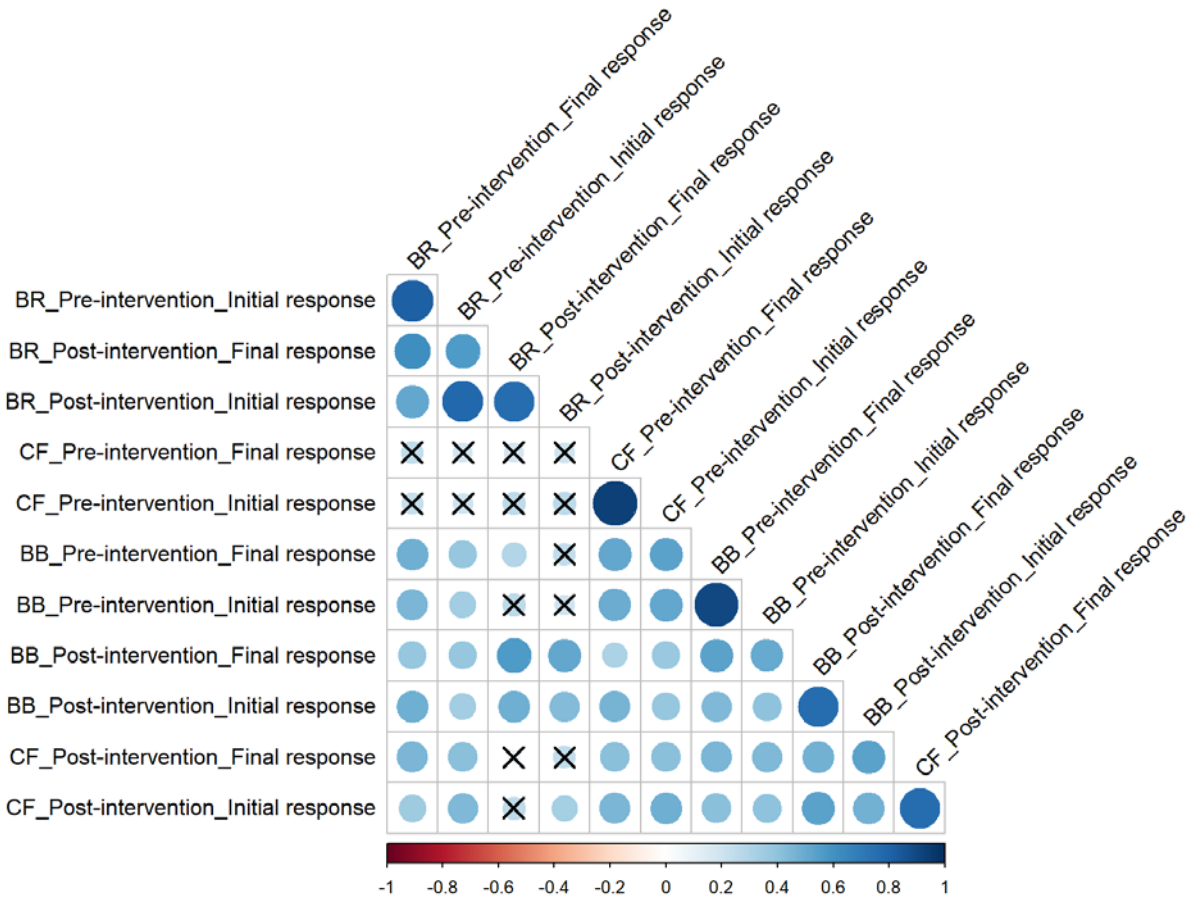


Figure S16. Correlogram in the training group in Session 3: For all task (BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks), blocks (pre- and post-intervention) and response stages (initial and final). Insignificant coefficients are marked with a cross.

Reliability index:

Table S17.

Cronbach's alpha for each task (BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks) and for compound scores (All = the composite mean of the three tasks) across the three training sessions.

Task	Initial response		Final response	
	Pre-intervention	Post-intervention	Pre-intervention	Post-intervention
BB	.82	.79	.85	.77
BR	.78	.81	.73	.73
CF	.78	.81	.81	.85
All	.79	.82	.75	.78