Debiasing in motion: Boosting sound intuiting through animated video training

Nina Franiatte*[1,2], Esther Boissin[3], 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

[3]Department of Psychology, Cornell University, Ithaca, USA

*Corresponding author:

Nina Franiatte

LaPsyDÉ (UMR CNRS 8240)

Sorbonne - Labo A. Binet

Université Paris Cité

46 rue Saint Jacques

75005 Paris

France

E-mail address: franiatte.nina@gmail.com

Abstract

Recent debiasing studies have shown that an explanation of the correct solution to a reasoning problem can often improve the performance of initially biased reasoners. However, most of these training studies have relied on text-based interventions. While effective, they may have limitations in reaching broader audiences. In the present study, we tested whether an animated video training on classic reasoning tasks can improve participants' reasoning accuracy. Specifically, we examined the nature of the effect of video-based training: Whether, like previous text-based interventions, it not only improves deliberate performance but also directly boosts intuitive performance. We conducted six studies on three classic reasoning tasks: The base-rate neglect, conjunction fallacy, and bat-and-ball tasks. 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 Studies 1, 2, and 3, participants received either video training, text training, or no training (control). In Studies 4, 5, and 6, participants of video and text groups were re-tested and received additional training two months later. Results of Studies 1, 2, and 3 indicated that animated video training is effective in boosting reasoning performance, as early as the initial intuitive stage. Results of Studies 4, 5, and 6 indicated that this effect persisted after two months. Overall, video training tended to outperform mere text training. These findings are consistent with the wider literature on multimedia research and can serve as a proof-of principle for a video debiasing training approach.

***Keywords*:** Reasoning · Dual-process theory · Debiasing · Video training

Introduction

Whenever there is a plane crash, it often gives rise to the belief that air travel is unsafe. Indeed, the intense emotional impact combined with the widespread media coverage can lead people to overestimate the likelihood of these rare events. Yet, the base rate – the extremely low probability of a crash given the high number of daily flights – shows that flying remains one of the safest modes of transportation. This phenomenon, also known as "base-rate neglect" (see Kahneman & Tversky, 1973), is one of the most paradigmatic examples of biased thinking. Instead of relying on logical or probabilistic evidence (e.g., "planes are about 80 times safer than cars when considering deaths per distance travelled"), people tend to focus on the most emotional or subjective one (e.g., the fear of crashing) which ultimately push them to believe that air travel is unsafe.

Decades of research in reasoning and decision-making have highlighted that people often fail to reason logically because they tend to rely on fast, intuitive judgments rather than on more demanding, deliberate ones (e.g., Evans, 2008; Kahneman, 2011; Stanovich & West, 2000). While intuitive processes can sometimes be useful, they may also conflict with traditional logical or probabilistic considerations (Kahneman & Frederick, 2005), leading people to provide heuristic responses (e.g., overestimating the frequency of crashes in the above example).

A famous explanation for this biased thinking has been given by the dual process theory, which describes 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). The use of heuristics-and-biases tasks, such as the base-rate example introduced above, has suggested that reasoners who manage to solve problems correctly would revise an initially generated intuitive response after completing deliberative processing. In other words, correct responding would require correction of an intuitive "System 1" response by the slower and more demanding "System 2" processing (e.g., Kahneman, 2011; Morewedge & Kahneman, 2010). However, because reasoners tend to minimize demanding computations, they will often apply intuitive processes by default without considering that the correct answer could be different (Evans & Stanovich, 2013; Kahneman, 2011; Kahneman & Frederick, 2005). Consequently, most reasoners remain biased.

Considerable scientific effort has been dedicated to improving decision-making (Johan, 2024; Lilienfeld et al., 2009; Milkman et al., 2009). Some debiasing training studies have been successful in this regard (e.g., Boissin et al., 2021, 2022, 2024; Claidière et al., 2017; Franiatte et al., 2024a, 2024b; Hoover & Healy, 2017; Trouche et al., 2014). Specifically, these training studies have shown that a short text which outlines the typical biased response and the correct solution strategy for a specific problem can improve subsequent reasoning performance on that same problem.

While these training results are promising, they also raise important questions about the nature of the training effect. That is, one possible explanation for the training effect is that it specifically improves deliberate thinking, allowing people to deliberate effectively (i.e., to use their "System 2") and correct their intuitively generated heuristic response. This hypothesis aligns with the "corrective" dual process view, which posits that the deliberate "System 2" primarily serves to correct the intuitive "System 1" (e.g., Kahneman, 2011; Pennycook et al., 2015b).

Alternatively, it is also theoretically possible that the training intervention directly affects intuitive thinking. That is, once reasoners grasp the solution, they may no longer generate an incorrect intuitive response but instead intuitively apply the correct solution strategy, without the need for a corrective "System 2" deliberation process. This hypothesis aligns with the "trained intuitor" view, which posits that "System 1" processes can be trained to give correct, fast, and effortless responses (Boissin et al., 2021, 2022, 2023a; Franiatte et al., 2024a, 2024b; see also Reyna et al., 2015).

Determining the nature of the training effect is crucial as it not only deepens our understanding of debiasing mechanisms but also has practical importance. Critically, if debiasing training can help people intuit correctly, this would have strong theoretical and applied implications. As Boissin et al. (2021) noted, while helping people deliberate more is laudable, they often lack the time, resources, or motivation to do so in their daily life. If debiasing training only helps them to deliberately correct erroneous intuitions, their effect may be limited. Hence, strengthening logical intuitions can be highly beneficial in this context (Boissin et al., 2021).

Recent evidence lends some credence to the "trained intuitor" view (Boissin et al., 2021, 2022; Franiatte et al., 2024a, 2024b; Purcell et al., 2022). To determine whether the training affected participants' intuitive and/or deliberate reasoning, these studies typically present heuristics-and-biases tasks using a two-response paradigm (Thompson et al., 2011). In this paradigm, participants are instructed to give two consecutive responses to a given problem. Initially, they are asked to provide the first response that comes to mind as quickly as possible. 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 response. To be maximally sure that participants do not deliberate during the initial stage, they are forced to give their initial response under time-pressure while performing a concurrent load task which burdens their cognitive resources (Bago & De Neys, 2019). Since "System 2" processes are often conceived as time and resource demanding, by restricting both, possible deliberation is minimized during the initial stage and participants are maximally forced to rely on intuitive processing. Two-response findings indicate that although most participants initially gave incorrect responses, a brief debiasing training intervention led them to provide correct responses to similar problems afterward. Critically, these responses are typically correct as early as the intuitive stage (e.g., Boissin et al., 2021,

2022). This sound intuiting training effect shows that after training, people do not need to engage in a costly deliberation process to give correct responses. Their intuitive responses are already correct.

However, most of the intuitive training studies to date have relied on text-based interventions (e.g., Boissin et al., 2021, 2022; Franiatte et al., 2024a, 2024b; Hoover & Healy, 2017). While these interventions have proven effective, they may be limited in their ability to reach a broader audience. Related work, such as Morewedge et al. (2015), has used video-based interventions, which could offer a more engaging and accessible alternative, especially when targeting the general public (e.g., Höffler & Leutner, 2007; Mayer, 2005; Schnotz & Rasch, 2005). Similarly, several studies have demonstrated that providing explicit instruction about cognitive biases through video-based materials leads to improved reasoning performance. These improvements have been observed both immediately following the instruction (e.g., Heijltjes et al., 2014, 2015) and after a delay of two weeks (e.g., Van Peppen et al., 2018). Videos provide a dynamic and appealing format that may increase motivation and participation (e.g., Downs, 2014; Berney & Bétrancourt, 2016). Moreover, they are easily scalable and can be distributed widely across various platforms, making them accessible to specific audiences (e.g., schools or workplaces; see Aalioui et al., 2022; Brown et al., 2007). This scalability offers a distinct advantage for debiasing efforts aimed at large-scale public engagement. However, previous studies using video formats have yet to test their effectiveness in improving intuitive reasoning.

Against this backdrop, decades of multimedia research have identified different functions that visual information can play relative to text-based content (e.g., Carney & Levin, 2002; Mayer, 2005). Empirical evidence strongly supports the idea that learners generally benefit more from a combination of visual and verbal information, than by text alone - a concept known as the “multimedia principle” (see Mayer, 2002, 2005). While the multimedia principle originally referred to text combined with illustrations, it has since expanded to include a range of coordinated visual and verbal formats (Butcher, 2014). In particular, numerous studies have recently explored how learning outcomes are affected by animated videos compared to text alone. The findings, however, are mixed. An earlier review by Tversky et al. (2002) found no consistent advantage of animations over static visuals or texts. In contrast, a more recent meta-analysis reported that animations can offer significant learning benefits when the dynamic content explicitly represents the “to-be-learned” information (Höffler & Leutner, 2007). Additionally, previous research has shown that videos can be effective training methods for teaching cognitive skills (e.g., Downs, 2014; Haferkamp et al., 2011) and for debiasing (Morewedge et al., 2015).

In the present work, we tested whether an animated video training on three classic reasoning tasks can improve participants’ reasoning accuracy. Specifically, we examined the nature of the effect of video-based training: Whether, like previous text-based interventions, it not only improves

deliberate performance but also directly boosts intuitive performance. To do so, we created video-based interventions using stop-motion animation and a voice-over narration.

To test the generalizability, we designed three distinct types of videos, each addressing a different reasoning task including different logico-mathematical principles and heuristics: The base-rate neglect (Kahneman & Tversky, 1973) in Study 1, the conjunction fallacy (Tversky & Kahneman, 1983) in Study 2, and the bat-and-ball tasks (Frederick, 2005) in Study 3. For each task, a full session lasted approximately 20 minutes and consisted of three different blocks: A pre-intervention, an intervention, and a post-intervention. Participants were randomly assigned to one of three groups: Animated video training, text training, or control (no training). During the intervention block, participants who received the training were introduced with a short video for the animated video training group, or with a short text for the text training group, that explained the rationale behind the task. Participants of the control group received no explanation.

For each task, the sound intuiting training effect was assessed two months later, during which participants of the text and video conditions were first retested and then received a second round of training. The aim was twofold: First, to determine whether the training effect persisted over time, and second, to evaluate whether an additional training session could further enhance reasoning performance.

Studies 1, 2 and 3

Method

Studies 1, 2, and 3 aimed to determine the nature of the effect of video-based training, and to compare this effect with that of text-based training. To do so, we used three different heuristics-and-biases tasks, i.e., the base-rate neglect (Kahneman & Tversky, 1973) in Study 1, the conjunction fallacy (Tversky & Kahneman, 1983) in Study 2, and the bat-and-ball tasks (Frederick, 2005) in Study 3. Each study was pre-registered separately. However, for ease of presentation and given the similarity of effects across tasks, we report the composite score that combined all tasks in the main text. The interested reader can also find all individual task-level analyses in the Supplementary Material.

Preregistration and data availability

The study design and research questions were preregistered separately for each task. Each study was preregistered on the AsPredicted website (https://aspredicted.org) and stored on the Open Science Framework. No specific analyses were preregistered. All data, materials, and analysis scripts are also available on the Open Science Framework (https://osf.io/5fuh9).

Participants

Participants were recruited online, using the Prolific Academic website (http://www.prolific.com). Only native English speakers from Canada, Australia, New Zealand, the USA, or the UK were allowed to take part in the study. The experiment took about 20 minutes and participants were paid £2 for their participation.

Study 1: Base-rate neglect. In total, 150 reasoners participated in this study (75 females, *M age* = 40.5 years, *SD* = 14.0). 50 participants were randomly assigned to the video group, 50 to the text group, and 50 to the control group. Among them, 2 participants had not completed secondary school, 53 had secondary school as their highest level of education, and 95 reported a university degree.

Study 2: Conjunction Fallacy. In total, 150 reasoners participated in this study (77 females, *M age* = 41.7 years, *SD* = 14.2). 49 participants were randomly assigned to the video group, 51 to the text group, and 50 to the control group. Among them, 2 participants had not completed secondary school, 57 had secondary school as their highest level of education, and 91 reported a university degree.

Study 3: Bat-and-ball. In total, 150 reasoners participated in this study (73 females, *M age* = 40.7 years, *SD* = 13.3). 51 participants were randomly assigned to the video group, 50 to the text group, and 49 to the control group. Among them, 4 participants had not completed secondary school, 45 had secondary school as their highest level of education, and 101 reported a university degree.

Our sample size decision was based on Boissin et al.'s (2021) original study who tested 100 participants divided into two groups. Since we have three groups in each study, we decided to test 150 participants per study, to maintain a similar number of participants per group (i.e., around 50). This allowed us to detect a medium training effect (*d = 0.5*) between the pre- and post-intervention blocks with a power of 90%.

Materials

The structure of Studies 1, 2, and 3 was similar. Each training study was composed of three blocks presented in the following order: A pre-intervention, an intervention, and a post-intervention block. Each pre- and post-intervention block contained four conflict problems and four no-conflict problems (see further). During the intervention, three more conflict problems were presented. After each problem, participants of the training groups received a short video or text explanation of the rationale behind the task, while participants of the control group received no explanation. In total, each participant had to solve 19 problems. All these problems are presented in the Supplementary Material Section A.

Pre- and post-intervention

Study 1: Base-rate neglect (BR). We presented problems taken from Bago and De Neys (2017). Participants always received a description of the composition of a sample (e.g., "This study contains I.T. technicians and boxers"), base rate information (e.g., "There were 995 I.T. technicians and 5 boxers") and a description that was designed to cue a stereotypical association (e.g. "This person is strong"). 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 format for a conflict problem:

"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?"

Note that we label 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 "creative" or "rich") 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).

To ensure that possible correct or incorrect responses did not originate from guessing, we also presented no-conflict control problems. In these problems, the description triggered a stereotypical trait of a member of the largest group. The heuristic intuition thus cued the correct response (e.g., "Person "F" is strong. There are 995 boxers and 5 I.T. technicians" in the above example). We presented four conflict and four no-conflict 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).

Study 2: Conjunction fallacy (CF). We presented problems taken from Andersson et al. (2020) and Boissin et al. (2022). All conjunction problems presented a short personality description of a character, consisting of their name (e.g., "Errin"), their age (e.g., "27"), their previous studies (e.g., "design") and their hobby/interest (e.g., "sewing"). 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 caregiver") and one option presented a conjunction of this unlikely and a likely characteristic (e.g., "A caregiver and a fashion enthusiast"). Two other filler options presented a characteristic that was very unlikely (e.g., "An astronaut") and a conjunction of two unlikely characteristics (e.g., "A caregiver and a genealogist"). The following illustrates the full format for a conflict problem:

"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?"

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 fashion enthusiast" 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 rare in our studies (i.e., 9.3% of options). However, strictly speaking, participants who select the singular very unlikely option (e.g., "An astronaut" in the above example) do not violate the critical conjunction rule. As Boissin et al. (2022) mentioned, given that we are interested in learning effects, selection of the very unlikely option can be considered a correct response. Hence, we considered

answers on which the conjunction fallacy is avoided (i.e., unlikely and very unlikely answers) as correct answers. Figure S1 in Supplementary Material Section B give a detailed overview of the selection frequency of each individual response option.

Study 3: Bat-and-ball (BB). We presented problems taken from Raoelison and De Neys (2019). They were modified versions of the original bat-and-ball problem (Frederick, 2005) which used quantities instead of prices (e.g., "On the shelves one can find 470 screws and screwdrivers. There are 400 more screws than screwdrivers. How many screwdrivers are there?"). 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. In the no-conflict control version, the heuristic intuition cued the correct response (e.g., "On the shelves one can find 560 screws and screwdrivers. There are 500 screws. How many screwdrivers are there on the shelves?"; see De Neys et al., 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. 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).

Counterbalancing. For each reasoning task, two sets of problems were created in which the conflict status of each problem (i.e., conflict vs no-conflict) 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. In each task, half of the participants were presented with the first set of problems while the other half was presented with the second set. Hence, 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 problems within each task was also randomized.

Two-response format. Problems were presented using a two-response paradigm (Thompson et al., 2011). That is, participants are asked to provide two consecutive responses on each problem: 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). 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 × 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). They received feedback as to whether they chose the correct or incorrect pattern. Based on previous pre-testing that indicated the time needed to read the preambles, move the mouse, and click on a response option, a time limit of 3 seconds was chosen for all base-rate problems. Likewise, the time limit was set to 5 seconds for all conjunction fallacy problems, and 8 seconds for all bat-and-ball problems. For the three tasks, previous pretesting established that the time limits imposed a stringent time-pressure that forced participants to respond significantly faster than in a traditional unconstrained, one-response test format (Bago & De Neys, 2017, 2019; Boissin et al., 2022). Note that the time limit and cognitive load were only applied during the initial response stage and not during the subsequent final stage in which participants were allowed to deliberate.

Justification. For exploratory purposes, after the last problem of the post-intervention block - which was always a conflict problem – in each study, 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", or "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 Boissin et al. (2022) for base-rate, Franiatte et al. (2024) for conjunction fallacy, and Bago and De Neys (2019) for bat-and-ball tasks. A justification was considered as correct when it explicitly mentioned the use of the base-rate (e.g., "Greater number of I.T. technicians to boxers. For every 1 boxer there are 199 I.T. technicians, so the odds are stacked against it being a boxer"), when it explicitly referred to the conjunction principle (e.g., "There are always more people who are simply caregiver than caregiver AND fashion enthusiast"), or the correct calculation for the bat-and-ball (e.g., "470 in total – 400 screws = 70 screwdrivers/2, the response is 35"). Other justifications, whether they mentioned an incorrect calculation or unspecified statement (e.g., "I used the same logic as in the explanations") were coded as incorrect.

Consistent with previous studies, results indicated that the majority of correct responses was also correctly justified after training (video group: 88 correct justifications out of 126 correct

responses, i.e., 70%, and text group: 71 correct justifications out of 116 correct responses, i.e., 61%). The interested reader can find details in Table S1 in Supplementary Material Section C. Note that the justification was untimed and retrospective.

Intervention block

During the intervention, participants had to solve three additional conflict problems (i.e., three base-rate items in Study 1, three conjunction fallacy items in Study 2, or three bat-and-ball items in Study 3), without any cognitive or time constraint.

In the text group, after each problem, participants were given a short text explanation of the typical biased response and the correct solution strategy. In the video group, after each problem, participants watched a short animated video with a voice-over narrated by a native U.S. English speaker. The voice-over was based on the exact same text as the text group. In total, participants in the text group were presented with three explanations, and participants in the video group watched three videos (approximately 2 minutes long).

For both the video and text groups, 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. No personal performance feedback (e.g., "Congratulations" 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).

We present below an example of a text explanation and a screenshot of a video explanation (Figure 1) for the base-rate task. The interested readers can find all text explanations in Supplementary Material Section A, and all video explanations on the Open Science Framework (https://osf.io/5fuh9). Note that participants in the control group simply solved three problems, without receiving any explanation.

Question:

"This study contains Hollywood celebrities and bakers. Person 'C' is rich. There are 5 Hollywood celebrities and 995 bakers. Is Person 'C' more likely to be a Hollywood celebrity or a baker?"

Text explanation:

"The correct answer to the previous problem is that person C is most likely a baker. Most people think the answer is a "Hollywood celebrity", but this answer is wrong.

Most people base their answer solely on the description (“Person C is rich”). If this were the only information given, this answer would be correct, as it is likely that there are more rich Hollywood celebrities in the world than rich bakers.

However, in the problem, you also got information about the specific number of bakers and Hollywood celebrities in the group from which person C got drawn. You were informed that person C was drawn randomly from a group with 995 bakers and only 5 Hollywood celebrities. Since there are so many more bakers in the group than Hollywood celebrities (almost 200 times more!), it becomes more likely that person C is a baker. After all, while Hollywood celebrities are generally wealthier than bakers, some bakers are rich.

If you combine this with the vastly larger number of bakers in the group, it will be more plausible that you’re dealing with a rich baker.”

Video explanation:

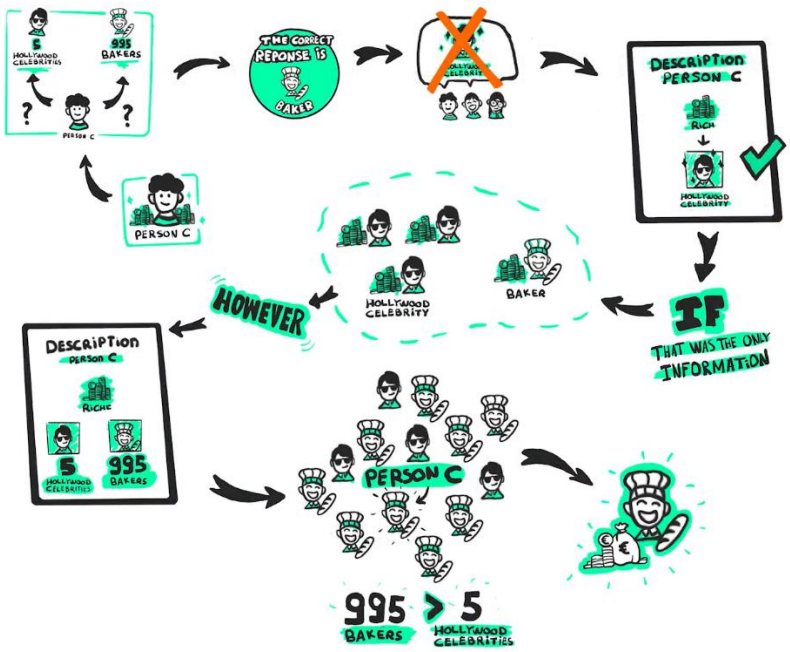

Figure 1. Screenshot of the explanation of a base-rate problem as presented in the animated video group.

Procedure

Debiasing training for base-rate, conjunction fallacy, and bat-and-ball tasks followed the same procedure. Each experiment was conducted online using the Qualtrics platform (https://www.qualtrics.com). First, participants were informed that the training would take around 20 minutes, require their full attention, and that they needed to have the sound on to follow the videos. A general description of the task was presented in which participants were instructed that they would

need to solve reasoning problems, 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 wanted to provide a final answer.

At the beginning of each task, participants solved two unrelated practice reasoning problems to familiarize themselves with the two-response procedure. 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., "Corey, 36, has previously studied journalism and likes gossip" for the conjunction fallacy 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 3000 ms (base-rate), 5000 ms (conjunction fallacy), or 8000 ms (bat-and-ball) to give their initial answer. Note that, in this initial "intuitive" response stage, the background of the screen turned yellow after 2000 ms (base-rate), 3000 ms (conjunction fallacy), or 6000 ms (bat-and-ball) to warn participants that they only had a short amount of time left to answer. If they had not provided an answer before the time limit, they were given a reminder that it was important to provide an answer within the time limit on subsequent trials (e.g., "You did not enter your response before the deadline. Try to respond within the deadline on the next 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. They received feedback as to whether their memory-response was correct. If the answer was not correct, they were reminded that it was important to perform well on the memory task on subsequent trials. Finally, the same reasoning problem was presented again, and participants were asked to provide a final "deliberate" answer (with no time limit nor cognitive load) and, once again, to indicate their confidence level.

At the end of each task, all participants were asked to complete a page with demographic questions. Additionally, participants in the video and text groups were asked to rate the clarity, enjoyment, and informativeness of the training intervention on a scale from 0 (not at all) to 10 (extremely). Note that, at the end of the study, participants of the control group were also presented with the explanations about how the base-rate, conjunction fallacy, or bat-and-ball problems could be solved.

Trial exclusion

Study 1: Base-rate neglect. Following our preregistration, we discarded trials in which participants failed to provide their initial answer before the deadline (i.e., 1.8%) or failed to pick the correct matrix in the load task (i.e., 13.4%). Therefore, we analysed the remaining 85.0% of all trials. On average, each participant contributed 14.1 (*SD* = 2.1) conflict trials out of 16, and 13.1 (*SD* = 2.8) no-conflict trials out of 16.

Study 2: Conjunction fallacy. Following our preregistration, we discarded trials in which participants failed to provide their initial answer before the deadline (i.e., 2.1%) or failed to pick the correct matrix in the load task (i.e., 13.2%). Therefore, we analysed the remaining 85.0% of all trials. On average, each participant contributed 13.6 (*SD* = 2.4) conflict trials out of 16, and 13.5 (*SD* = 2.3) no-conflict trials out of 16.

Study 3: Bat-and-ball. Following our preregistration, we discarded trials in which participants failed to provide their initial answer before the deadline (i.e., 1.1%) or failed to pick the correct matrix in the load task (i.e., 14.2%). Therefore, we analysed the remaining 84.9% of all trials. On average, each participant contributed 13.1 (*SD* = 2.4) conflict trials out of 16, and 14.0 (*SD* = 2.3) no-conflict trials out of 16.

Note that, for each task, the number of excluded trials was highly similar across the three groups (i.e., video, text, and control groups). Detailed results of trial exclusions for each task and group are provided in Table S10 in Supplementary Section J.

Composite measure

We preregistered three separate studies (each corresponding to a different task, i.e., base-rate neglect, conjunction fallacy, and bat-and-ball). However, as the results were highly consistent across tasks, we combined them for ease of presentation. Specifically, we calculated a score by averaging the proportion of correct initial and final responses for each participant and each task. We then combined these scores into a single composite variable. The figures and tables in the main text present both the combined and individual data. For completeness, we also calculated the composite score for no-conflict trials (see Table S3 in Supplementary Material Section D).

Statistical analyses

The data were processed and analysed using the R software (R CoreTeam, 2017) and the following packages (in alphabetical order): ggplot2 (Wickham, 2016), lmerTest (Kuznetsova et al., 2017) and tidyverse (Wickam et al., 2024). Throughout the article, we used mixed-effect regression models in which participants were entered as random intercepts. The Wald test was used to assess the

statistical significance of the model's fixed effect. Note that we tried to design a more complex model, including both participants and items as random intercepts, but it failed to converge. Thus, we kept the simpler model described above.

Results

Conflict trials accuracy

First, we tested whether the training improves reasoners' performance after the intervention, and whether it specifically boosts intuitive or deliberative responses. To do so, we focused on the average proportion of correct initial and final responses on conflict trials, in each block (pre- and post-intervention) and in each group (video, text, and control), and their interaction.

Figure 2 shows the final conflict trials accuracies and indicates that, on average, participants either from the training groups or the control group had comparable performance before the intervention. Indeed, a majority of them was biased and gave incorrect responses even when allowed to deliberate (video group: *M* = 31.7%, *SD* = 43.1; text group: *M* = 32.2%, *SD* = 41.0; control group: *M* = 30.0%, *SD* = 41.6). After the training, the average proportion of correct final responses improved sharply in the video and text training groups (respectively, 49 points rise, reaching *M* = 80.9%, *SD* = 34.6, and 40 points rise, reaching *M* = 72.3%, *SD* = 38.6), while the control group's improvement was marginal (6 points rise, reaching *M* = 36.4%, *SD* = 44.8). The Block x Group interaction was significant, $\chi2$ (2) = 92.4, $p < .001$.

The same tendencies were observed for initial responses. Participants showed low initial accuracies before the intervention (video group: *M* = 23.9%, *SD* = 36.7; text group: *M* = 25.3%, *SD* = 36.4; control group: *M* = 22.4%, *SD* = 35.5). Performance significantly increased in video and text groups after the intervention (respectively, 53 points rise, reaching *M* = 76.6%, *SD* = 35.2, and 42 points rise, reaching *M* = 67.7%, *SD* = 40.1), while it was less pronounced for the control group (11 points rise, reaching *M* = 33.3%, *SD* = 42.2). The Block x Group interaction was significant, $\chi2$ (2) = 93.4, $p < .001$. This suggests that video and text training interventions improve both initial intuitive and final deliberate performance. Note that these tendencies were also observed on each individual task (see Figure 2, bottom panels).

Second, we tested whether the effect of the text training on performance differs from the control group. To do so, we focused on the average proportion of correct initial and final responses on conflict trials, in each block (pre- and post-intervention) and in each group (text vs control), and their interaction. The Block x Group interaction was significant both for the final: $\chi2$ (1) = 63.4, $p < .001$, and the initial: $\chi2$ (1) = 54.9, $p < .001$, performance. These results are consistent with previous training

studies which showed that a short text training can boost sound reasoning (e.g., Boissin et al., 2021, 2022, 2024; Franiatte et al., 2024a, 2024b).

Then we tested whether the effect of the video training on performance differs from that of the control group and whether it is comparable to the effect of the text training. To do so, we focused on the average proportion of correct initial and final responses on conflict trials, in each block (pre- and post-intervention) and in each group (video vs control, and video vs text), and their interaction. Regarding the difference between the video training and the control no-training groups, results revealed a significant Block x Group interaction for the final: $\chi2$ (1) = 91.9, $p < .001$, and initial: $\chi2$ (1) = 96.1, $p < .001$ responses. This indicates that a short video training can boost correct reasoning performance at both intuitive and deliberate levels, beyond the spontaneous increase that is observed in the control group. Finally, the comparison between video and text groups indicated a marginal Block x Group interaction for final responses: $\chi2$ (1) = 3.4, $p = .064$, and a significant interaction for initial responses: $\chi2$ (1) = 4.5, $p = .034$, with video intervention showing a slightly better training effect than text intervention.

No-conflict trials accuracy

For completeness, we also analysed the average proportion of initial "intuitive" and final "deliberate" correct responses on all no-conflict problems.

As expected, performance was consistently near ceiling in pre-and post-intervention blocks for both final responses (*M* = 91.5%, *SD* = 21.0 in the video group, *M* = 93.4%, *SD* = 16.0 in the text group, and *M* = 89.0%, *SD* = 23.6 in the control group) and initial responses (*M* = 89.2%, *SD* = 22.9 in the video group, *M* = 91.3%, *SD* = 19.1 in the text group, and *M* = 86.4%, *SD* = 25.0 in the control group). In line with previous studies (e.g., Bago & De Neys, 2020), participants' high accuracy rates on the no-conflict problems suggested that they were paying attention to the task and avoided random guessing. It also helps dismiss a possible alternative explanation for the training effect. One could argue that the intervention simply cued a "reversed" heuristic. That is, participants would deduce that they are being presented with counter-intuitive trick problems in which the right answer is always the opposite of the cued heuristic/stereotypical response (e.g., "Select the opposite of what you believe to be the correct answer", see Boissin et al., 2022). This would lead to selection of the correct response on conflict problems. However, such a "reversed heuristic" strategy would have led to a floored post-intervention performance on the no-conflict problems (in which the intuitive, heuristic response was always correct). Hence, the consistent high accuracies on our no-conflict (control) problems argue against this (see Table S3 in Supplementary Material Section D for full results).

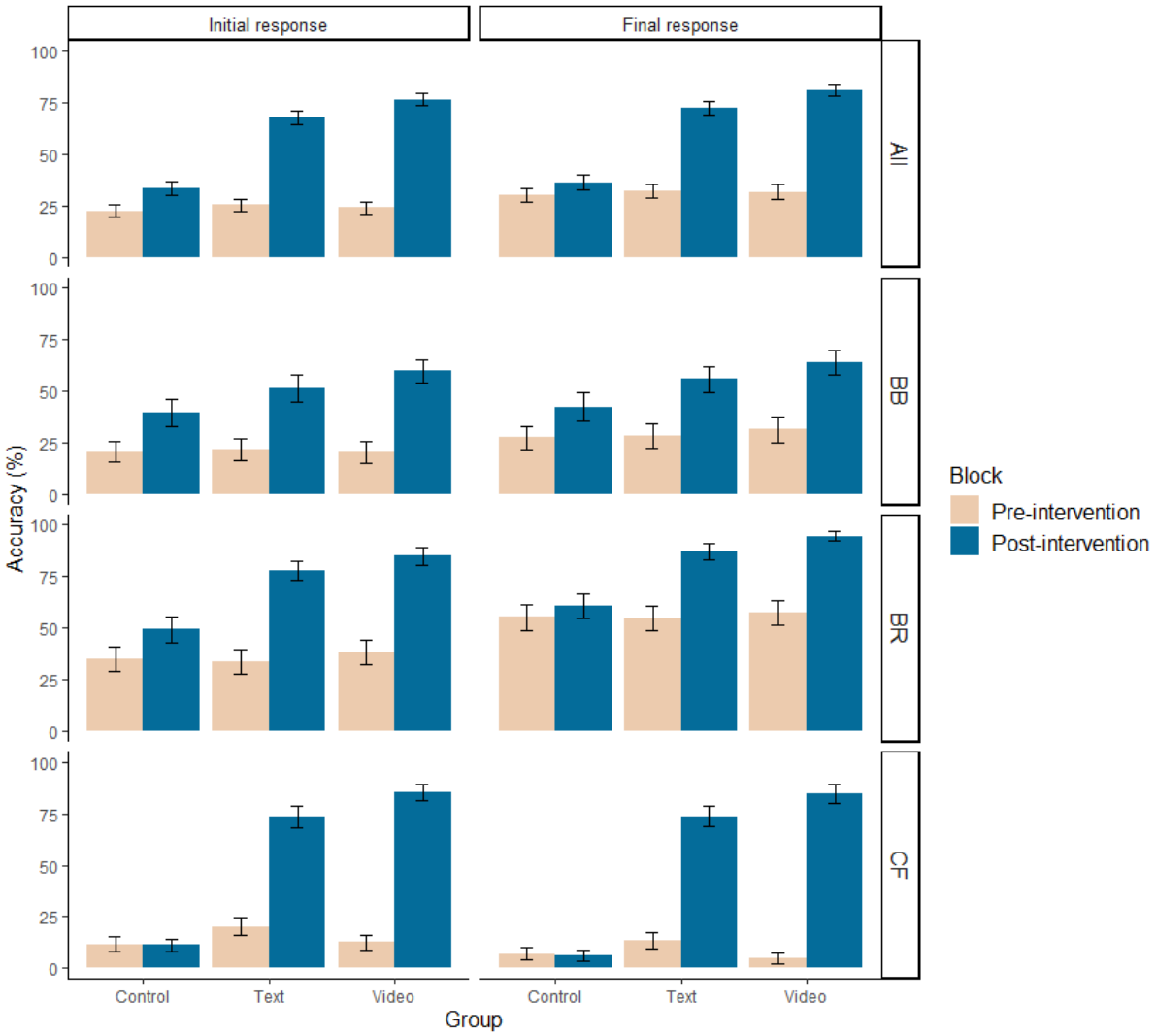

Figure 2. Mean accuracy (%) of correct initial and final responses on conflict problems for control, text, and video groups, before and after the intervention, for each task (BB, BR, CF), and combined (All). Error bars are standard errors. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy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

Direction of change

To gain some deeper insight into how people changed (or did not change) their response after deliberation, we conducted a direction of change analysis on conflict problems (Bago & De Neys, 2017, 2019). Specifically, each trial is composed of two responses, the initial "intuitive" one (given under time pressure and cognitive load) and the final "deliberate" one. Correct responses are labelled "1" and incorrect responses are labelled "0". Hence, each trial can result in one of four different patterns: "00" pattern (incorrect response at both response stages), "11" pattern (correct response at both response stages), "01" pattern (initial incorrect and final correct responses), and "10" pattern (initial correct and final incorrect responses). Figure 3 plots the direction of change distribution for each group, in pre- and post-intervention blocks.

Consistent with the overall accuracies presented above, a large number of conflict trials had a "00" (biased) pattern before the intervention (video group: *M* = 63.5%, *SD* = 43.2; text group: *M* = 61.3%, *SD* = 42.5; control group: *M* = 66.9%, *SD* = 41.2). Following the intervention, trials which produced "00" patterns reduced in all groups, with a bigger decrease for the video and text training groups (respectively, 48.4 and 38.7 points drop) compared to the control group (7.3 points drop). This decrease in "00" patterns after the interventions led to an increase in "11" patterns, that was more pronounced for the video (53.8 points rise) and the text training groups (43.8 points rise) than for the control group (9.7 points rise). It is also important to note that this decrease in "00" patterns did not lead to an increase in "01" patterns (4.3 points drop for the video group, 3.7 points drop for the text group, and 3.7 points drop for the control group). This suggests that video and text training interventions mainly helped participants intuit the correct solution strategy rather than correct an initial "erroneous" response through deliberation. Hence, we replicated previous debiasing studies showing that a short text training can help people intuit correctly (e.g., Boissin et al., 2021, 2022; Franiatte et al., 2024a, 2024b). Critically, these results also suggest that a short video training can boost sound intuiting. Note that similar trends were observed for each individual reasoning task (see Figure 3, bottom panels).

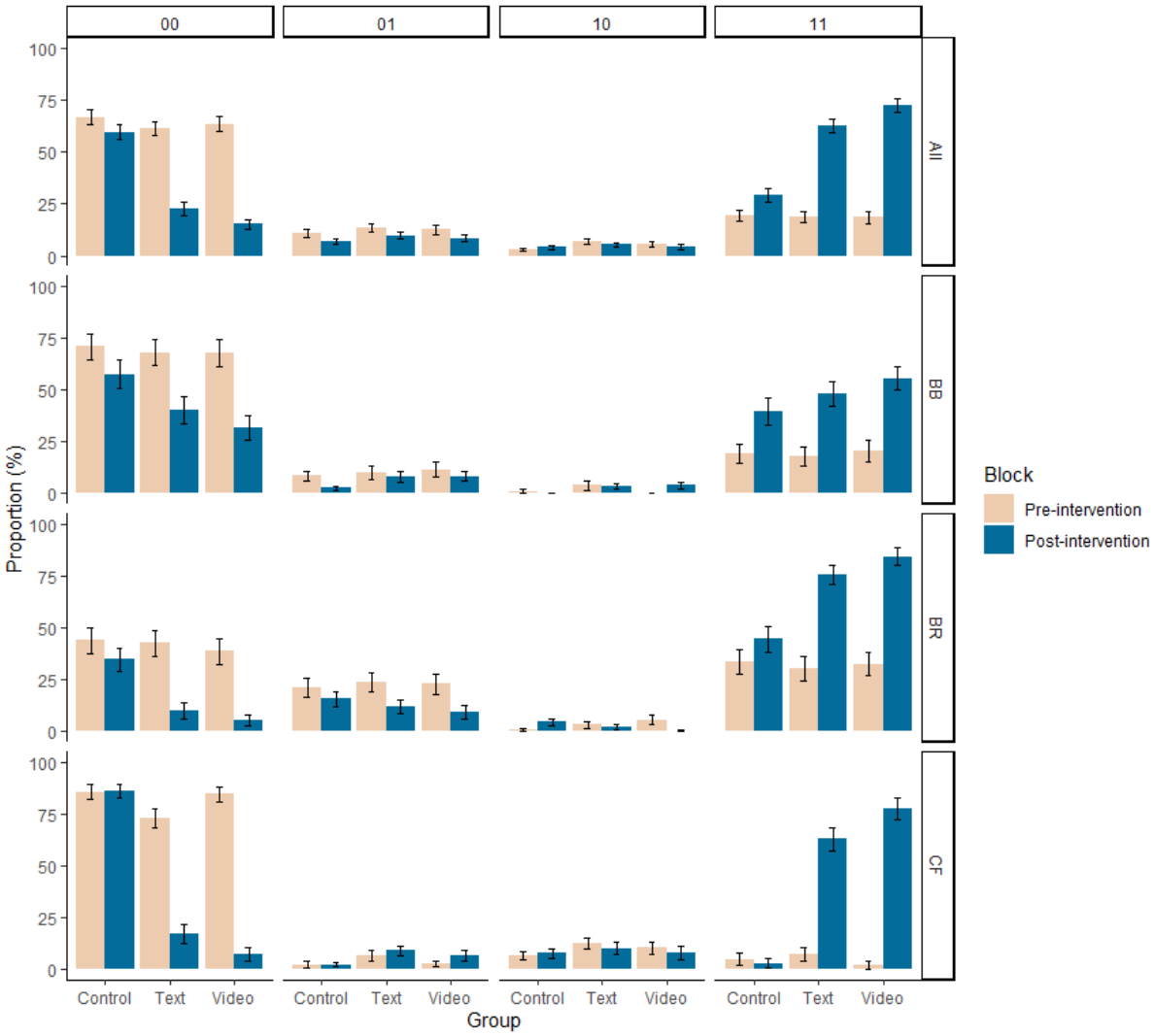

Figure 3. Proportion (%) of each direction of change (i.e., "00" pattern, "01" pattern, "10" pattern, and "11" pattern; 0 = incorrect response, 1 = correct response, first digit = initial response, second digit = final response) on conflict problems, before and after the intervention, for each task (BB,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

Individual level direction of change

To gain some deeper insight into how a given reasoner changed (or did not change) their response, we also performed an individual level accuracy analysis on conflict problems (Raoelison & De Neys, 2019). Thus, for each participant, we focused on their dominant direction of change before and after the intervention and classified it using the categories introduced by Boissin et al. (2021, 2022). As before, we present the results for the three reasoning tasks combined (i.e., at the composite level). Analyses for each individual task can be found in Supplementary Material Section E.

First, participants who predominantly provided incorrect responses (i.e., "00" patterns) before and after the intervention are labelled as "biased" responders. They represented 15.8% in the video

group, 25.7% in the text group, and 61.1% in the control group. Second, participants who provided a stable majority of correct answers ("01" or "11" trials) before and after the intervention are labelled as "correct" responders. They represented 21.2% in the video group, 22.3% in the text group, and 22.2% in the control group. Third, participants whose accuracy increased after the intervention are labelled "improved" responders. They either gave a majority of biased responses ("00" patterns) before the intervention and then switched to a majority of correct responses after the intervention ("01" or "11" patterns), or already gave a majority of correct final responses ("01" patterns) before the intervention but switched to a majority of correct initial and final responses ("11" patterns) after the intervention. They amounted to 62.3% in the video group, 52.0% in the text group, and 12.8% in the control group. Participants who gave inconsistent response patterns and could not be classified were put in the "other" category (1.0% in the video group, and 4.0% in the control group).

Additional analyses

Conflict detection. Previous work in the reasoning field showed that despite giving an incorrect response, reasoners often show some conflict or error sensitivity - as expressed for example in decreased confidence in their erroneous conflict trial responses (see De Neys, 2022 for review). In the present work, we explored whether video and text interventions 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.

For each problem, participants were asked to rate their confidence in the correctness of their answer after responding, on a scale from 0%, absolutely not confident, to 100%, absolutely confident. We used the typical conflict detection index introduced in the study of De Neys et al. (2011), by contrasting confidence ratings for correctly solved no-conflict problems to confidence ratings for incorrectly solved conflict problems. We compared this index before and after the intervention, in each of the three groups (i.e., video, text, and control). Note that a higher index is assumed to reflect a more pronounced conflict or error detection sensitivity. Following our preregistrations, we focused on initial response conflict detection since it gives a purer measure of intuitively experienced conflict (e.g., see Voudouri et al., 2022). Overall, effects were small and were not consistent across tasks: While the conflict detection index slightly improved after training in the video group across all three tasks, improvements in the text group were only observed for the base-rate and conjunction fallacy tasks. By and large, the training intervention did not seem to significantly enhance biased reasoners' ability to detect conflict. The interested reader can find full results in Supplementary Material Section F.

Predictive conflict detection. We also used confidence ratings to test the predictive effect of conflict detection, i.e., to determine whether one's ability to detect conflict before the intervention could

predict a better success of the training intervention. That is, we analysed whether reasoners who improved their performance after the video or text 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 in the text and video groups, before the intervention. Overall, here results pointed towards a consistent better conflict detection among improved than biased reasoners, both in the video and text groups (video group: $M_{improved}$ = 13.9%, $SD_{improved}$ = 24.0, and M_{biased} = 5.3%, SD_{biased} = 17.0; text group: $M_{improved}$ = 11.5%, $SD_{improved}$ = 21.0, and M_{biased} = 1.2%, SD_{biased} = 10.9; see Supplementary Material Section G for details). Hence, in line with previous findings with text-based training only (Boissin et al., 2021, 2022; Franiatte et al., 2024a, 2024b), reasoners who started to respond correctly after the intervention (i.e., improved ones) seem to be characterized by more pronounced conflict detection before the intervention

Ratings. Our results so far indicate that video training is effective in debiasing individuals for both intuitive and deliberative responses, with a slightly stronger effect compared to text-based training. One potential explanation for this may lie in the format itself. The video format could be more engaging and motivating than text, making it more appealing and potentially encouraging higher engagement during the intervention. To test this hypothesis, we analysed participants' preferences for text or video training. At the end of each task, participants in the text and video groups were asked to rate on a scale from 0 (not at all) to 10 (extremely) the clarity, enjoyment, and informativeness of the explanations they received (see 2.1.6 Procedure). By and large, we observed high ratings in both the video and text groups (all scales average > 6.5) with no clear differences between the groups (see Supplementary Material Section H for all results).

Studies 4, 5 and 6

Studies 1, 2, and 3 showed that a short video debiasing training can help people reason more accurately, as early as the intuitive stage. Two months after completing the first training session, participants in the video and text groups were invited in a retest followed by a second training session. Our objectives were threefold: First, to test whether the effect of the video training was robust and sustained over time; second, to determine whether a second training session could further enhance performance; and third, to compare these effects with those of the text-based training. For each task,

after a new pre-intervention test (that served as a re-test), participants again went through our video or text intervention and completed a final post-intervention block.

Method

We again ran three independent studies (one for each of the three reasoning tasks). As before, given the consistency of results across tasks, we again combined the analyses into a single composite score. For clarity in presentation, we report the aggregated results in the main text. As before, the figures and tables in the main text present both combined and individual data. The interested reader can also find all individual task level analyses in the Supplementary Material.

Preregistration and data availability

We preregistered the study design and research questions separately for each task (i.e., base-rate neglect, conjunction fallacy, and bat-and-ball). Each of the three study was preregistered on the AsPredicted website (https://aspredicted.org) and stored on the Open Science Framework. No specific analyses were preregistered. All data, material, and analysis scripts are also available on the Open Science Framework (https://osf.io/5fuh9).

Participants

All participants from the training groups who completed the first training session were contacted again and invited to participate. The experiment took about 20 minutes and participants were paid £3 for their participation. Note that there was no control group. For ethical reasons, control group participants were given the training explanations at the end of the first training session. Consequently, they could no longer serve as a no-training control group for the retest. We accepted participations upon two weeks after launching each second training session.

Study 1: Base-rate neglect. All 100 participants from video and text groups who completed the first training session were contacted again and invited to participate. In total, 73 participants (i.e., 73%) took part in the re-test (40 females, *M age* = 42.6 years, *SD* = 13.2). The sample was composed of 36 participants in the video group, and 37 in the text group.

Study 2: Conjunction Fallacy. All 100 participants from video and text groups who completed the first training session were contacted again and invited to participate. In total, 75 participants (i.e., 75%) took part in the re-test (35 females, *M age* = 43.7 years, *SD* = 15.2). The sample was composed of 35 participants in the video group, and 40 in the text group.

Study 3: Bat-and-ball. All 101 participants from video and text groups who completed the first training session were contacted again and invited to participate. In total, 77 participants (i.e., 76%) took part in the re-test (34 females, *M age* = 44.8 years, *SD* = 14.6). The sample was composed of 39 participants in the video group, and 38 in the text group.

Materials and procedure

Two months after the first training session, participants were invited to a second training session. The procedure followed the same structure as the initial session, with the only difference being the use of different content materials between the sessions (see Supplementary Material Section A; see Table S2 in Supplementary Material Section C for justification data).

Trial exclusion

Study 1: Base-rate neglect. Following our preregistration, we discarded trials in which participants failed to provide their initial answer before the deadline (i.e., 2.4%) or failed to pick the correct matrix in the load task (i.e., 10.7%). Therefore, we analysed the remaining 87.2% of all trials. On average, each participant contributed 14.4 (*SD* = 1.7) conflict trials out of 16, and 13.4 (*SD* = 2.2) no-conflict trials out of 16.

Study 2: Conjunction fallacy. Following our preregistration, we discarded trials in which participants failed to provide their initial answer before the deadline (i.e., 3.0%) or failed to pick the correct matrix in the load task (i.e., 10.7%). Therefore, we analysed the remaining 86.6% of all trials. On average, each participant contributed 13.8 (*SD* = 2.2) conflict trials out of 16, and 13.9 (*SD* = 2.0) no-conflict trials out of 16.

Study 3: Bat-and-ball. Following our preregistration, we discarded trials in which participants failed to provide their initial answer before the deadline (i.e., 1.5%) or failed to pick the correct matrix in the load task (i.e., 11.6%). Therefore, we analysed the remaining 87.1% of all trials. On average, each participant contributed 13.8 (*SD* = 2.7) conflict trials out of 16, and 14.0 (*SD* = 2.6) no-conflict trials out of 16.

We analysed the number of excluded trials per task in the two re-tested groups (i.e., the video and text groups). For each task, the number of excluded trials was highly similar across groups, and detailed results are provided in Table S11 in Supplementary Section J.

Composite measure

Consistent with the initial training session, we present a composite measure combining the results of the three tasks. We first averaged the proportion of correct initial and final responses for each participant and each task, and then combined these scores into a single composite variable. For completeness, we also calculated the composite performance for no-conflict problems (see Table S4 in Supplementary Material Section D).

Results

The sustained training effect

To test whether the training effect sustained over time, we compared performance on conflict items for video and text groups between the post-intervention block of the initial session (i.e., after the first training) and the pre-intervention block of the second training session (i.e., "retest" two months later). We also tested whether performance in the pre-intervention block of the second training session was higher than that in the pre-intervention block of the initial training session.

Conflict trials accuracy. First, we focus on final response accuracies. Figure 4 shows that performance slightly decreased after two months. Participants in the video and text groups tended to provide fewer correct responses two months after the first training (respectively, M = 62.4%, SD = 45.4, and M = 46.6%, SD = 46.4 in the pre-intervention block of the second training session) compared to immediately after it (respectively, M = 80.9%, SD = 34.6, and M = 72.3%, SD = 38.6 in the post-intervention of the first training session). In other words, after two months, performance dropped by 18.5 points in the video group, $t(196) = 3.58$, $p < .001$, $d = .46$, and by 25.7 points in the text group, $t(217) = 4.80$, $p < .001$, $d = .60$. Nevertheless, reasoners still gave more correct final responses two months after training (in the pre-intervention block of the second training session; M = 62.4%, SD = 45.4 in the video group, and M = 46.6%, SD = 46.4 in the text group) than before their first training (in the pre-intervention block of the first training session; M = 31.7%, SD = 43.1, $t(228) = 5.49$, $p < .001$, $d = 0.69$ in the video group, and M = 32.2%, SD = 41.0, $t(226) = 2.63$, $p = .009$, $d = 0.33$ in the text group). Hence, despite a decrease in performance, participants of both video and text groups provided more correct responses after two months than before the first training, indicating that the training effect on final responses sustained after two months.

In the same vein, focusing on initial responses, Figure 4 shows that participants from the video and text groups gave less correct responses two months after the initial training session (respectively, M = 51.8%, SD = 42.4, and M = 39.7%, SD = 43.3) than just after it (respectively, M = 76.6%, SD = 35.2, and M = 67.7%, SD = 40.1). This corresponds to a drop of 24.8 points after two months in the video

group, $t(208) = 5.0$, $p < .001$, $d = 0.64$, and a drop of 28.0 points after two months in the text group, $t(233) = 5.37$, $p < .001$, $d = 0.67$. Importantly, reasoners still gave more correct initial responses two months after the first training (in the pre-intervention block of the second training session; $M = 51.8\%$, $SD = 42.4$ in the video group, and $M = 39.7\%$, $SD = 43.3$ in the text group) than before their first training session (in the pre-intervention block of the first training session; $M = 23.9\%$, $SD = 36.7$, $t(215) = 5.53$, $p < .001$, $d = 0.70$ in the video group, and $M = 25.3\%$, $SD = 36.4$, $t(219) = 2.86$, $p = .005$, $d = 0.36$ in the text group). Hence, for initial responses, we also observed a slight decrease in performance after two months, but reasoners still had a significantly higher rate of correct responses compared to before the first training. In sum, the training effect also sustained after two months for initial responses.

To sum up, even if the text and video training effect diminishes after two months, performance still remained higher than before the first training. This suggests that the training effect is robust and sustained for at least two months, for the initial "intuitive" and final "deliberate" responses. These results were also backed up by a direction of change analysis (see Figure S9 in Supplementary Material Section I). Note that these tendencies were also observed on each individual task (see Figure 4, bottom panels).

The above analyses indicated that we observed a sustained training effect in both video and text groups. Indeed, initial and final accuracies were still higher two months after training than before the first training. As Figure 4 indicates, this effect also tended to be slightly stronger in the video group (final responses: +31 points, initial responses: +28 points above pre-training) than in the text group (final responses: +14 points, initial responses: +14 points above pre-training). Statistical analyses showed that these differences between video and text groups were significant (final responses: $t(216) = 2.89$, $p = .004$, initial responses: $t(219) = 2.15$, $p = .03$). The performance drop from post first training to the pre-intervention level two months later was also slightly less pronounced in the video group (final responses: -25 points, initial responses: -19 points) than in the text group (final responses: -28 points, initial responses: -26 points). To test this statistically we compared initial and final performance of the two training groups (i.e., video and text) after two months (i.e., between post first training to the pre-intervention level two months late). However, this difference did not reach significance (final responses: $t(220) = 1.19$, $p = .24$, initial responses: $t(220) = 0.01$, $p = .99$).

In the second training session, we managed to reach 75% (225/301) of the first session participants (i.e., video and text groups). To check for a possible attrition confound (e.g., subjects who did better in the first session were more likely to sign-up for the second session), we compared the first session pre-intervention conflict problem accuracy of the subgroup of the second session participants (video group: initial responses: $M = 23.9\%$, $SD = 36.1$, final responses: $M = 31.3\%$, $SD = 43.4$; text group: initial responses: $M = 25.3\%$, $SD = 36.9$, final responses: $M = 32.9\%$, $SD = 40.7$) to the accuracy of the first session pre-intervention of participants in video and text groups who did not take

part - but were invited - to the re-test (video group: initial responses: *M* = 23.9%, *SD* = 39.0, final responses: *M* = 32.7%, *SD* = 42.5; text group: initial responses: *M* = 25.2%, *SD* = 35.5, final responses: *M* = 29.9%, *SD* = 42.4). Given that both groups showed very similar accuracy rates (video group: initial responses: $t(63) = .0007$, $p = 1.0$, $d = 0.0001$, final responses: $t(68) = .18$, $p = .86$, $d = 0.03$; text group: initial responses: $t(61) = .009$, $p = .99$, $d = 0.002$, final responses: $t(57) = .38$, $p = .71$, $d = 0.07$), it is unlikely that the second training session results are artificially boosted because of an attrition confound.

No conflict trials accuracy. For completeness, note that no-conflict problem accuracies were also analysed. Performance was consistently near ceiling in pre- and post-intervention blocks for initial and final responses (see Table S4 in Supplementary Material Section D).

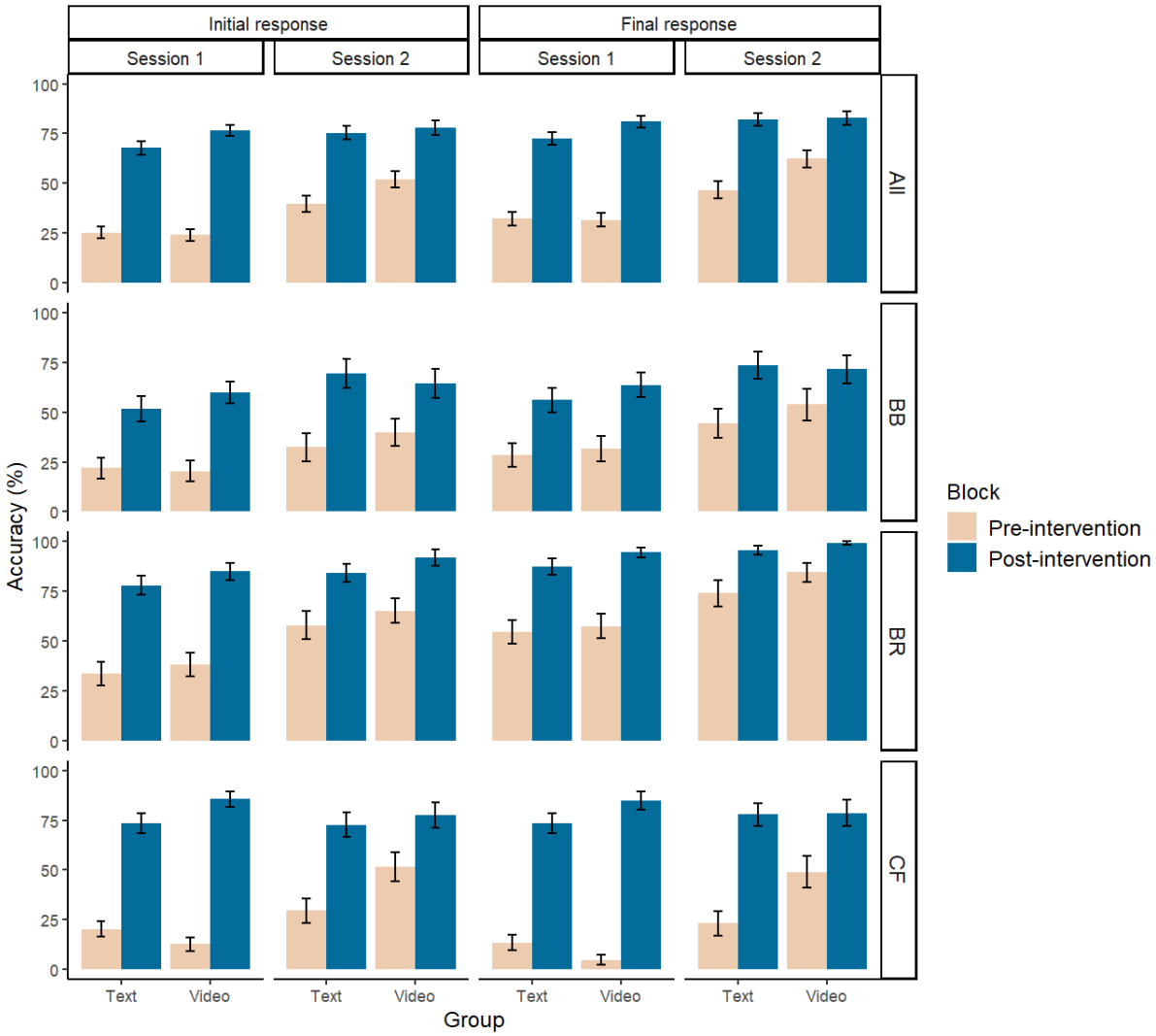

Figure 4. Mean accuracy (%) of correct initial and final responses on conflict problems for text and video groups, before and after the first and the second training sessions, for each task (BB, BR, CF), and combined (All). Error bars are standard errors. BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy tasks, All = the composite mean across the three tasks.

The second training effect

We also tested whether a second training session could further improve reasoning performance. Consequently, we compared conflict accuracies across the pre-retest and post-intervention blocks of the second training session, and across the post-intervention blocks of the first and second training sessions.

Conflict trials accuracy. With respect to the comparison of conflict accuracies across the pre-retest and post-intervention blocks of the second training session, as Figure 4 shows, it is clear that the second training again boosted final performance both for the video group (+20.5 points, reaching M = 82.8%, SD = 35.9 in the post-intervention block of the second session, $t(207) = 3.71$, $p < .001$, $d = 0.50$)

and the text group (+35.5 points, reaching *M* = 82.2%, *SD* = 34.1, in the post-intervention block of the second session, $t(207) = 6.62$, $p < .001$, $d = 0.88$).

Participants also gave more correct initial responses after the second training. In the video group, performance rose by 25.5 points (reaching *M* = 77.8%, *SD* = 38.1 in the post-intervention block of the second session, $t(215) = 4.76$, $p < .001$, $d = 0.64$). Similarly, in the text group, performance rose by 36.3 points (reaching *M* = 75.4%, *SD* = 37.7 in the post-intervention block of the second session, $t(222) = 6.64$, $p < .001$, $d = 0.88$). Hence, the slight performance decrease two months after the initial training was completely remediated with an additional training. Given that the training effect in the video group tended to be more sustained, participants in the text group also tended to show greater retraining benefits (final responses: $t(222) = 2.81$, $p = .005$, initial responses: $t(222) = 2.07$, $p = .04$).

Additionally, we also tested whether a second training session could further improve reasoning performance above the first training by comparing conflict accuracies across the post-intervention blocks of the first and second training sessions. Figure 4 shows that although performance after the first training was already high, we found a slightly better post-intervention accuracy after the second training both for initial and final accuracies. However, this difference reaches significance only for the increase in final responses after the second training in the text group, $t(258) = 2.21$, $p = .028$, $d = 0.27$. Indeed, this was not the case for final responses in the video group, $t(230) = 0.43$, $p = .67$, $d = 0.05$, nor for initial responses in the text group, $t(253) = 1.61$, $p = .11$, $d = 0.20$, or in the video group, $t(222) = 0.25$, $p = .81$, $d = 0.03$.

These results were also backed up by a direction of change analysis (see Figure S9 in Supplementary Material Section I).

No-conflict trials accuracy. For completeness, no-conflict problem accuracies were also analysed. As in the first training session, performance was consistently near ceiling in pre- and post-intervention blocks for initial and final responses (see Table S4 in Supplementary Material Section D).

Individual level direction of change. We performed an individual level accuracy analysis using the four categories (“correct”, “biased”, “improved”, “other”) defined in the first training session. Reflecting the overall accuracy effects, throughout the second training, a higher number of reasoners were already labelled as correct in the video group (54.1%) than in the text group (38.9%). Relatedly, a smaller number of reasoners improved in the video group (28.4%) compared to the text group (45.1%). These tendencies led to a similar number of biased reasoners in the two groups at the end of the second training session (video: 15.6%, text: 15.9%). Participants who gave inconsistent response

patterns and could not be classified were put in the "other" category (1.8% in the video group; see Supplementary Material Section E for full results).

Additional analyses. For completeness, although there were few biased reasoners remaining, we also looked at conflict and predictive conflict detection for the second training. We did not identify clear systematic trends (see Supplementary Material Sections F and G for full results)**.**

Discussion

The present study aimed to investigate whether an animated video debiasing training can improve participants' reasoning accuracy on three reasoning tasks (i.e., base-rate neglect, conjunction fallacy, and bat-and-ball). Specifically, we examined the nature of the video training effect: Whether it improves deliberate and/or intuitive reasoning performance, and whether its effect is comparable to that of text-based debiasing training. In an initial session, participants received either video training, text training, or no training (control). Two months later, during a second session, participants in the video and text groups were first retested and then received a second round of training. We used a two-response paradigm to track participants' initial "intuitive" and final "deliberate" responses.

Results showed that animated video training proved to be effective at boosting reasoning performance. Notably, this improvement occurred as early as the initial "intuitive" stage, resulting in an overall 53% increase in performance (vs 42% in the text group). This suggests that, in line with previous work with text-based training, after animated video training, reasoners were typically able to favour the correct response over a biasing stereotypical belief (for base-rate and conjunction fallacy) or a conflicting cued heuristic mathematical response (for bat-and-ball) without further need for deliberation. This animated video training even tends to (slightly) outperform a purely text-based training.

Critically, the second video training session revealed that debiasing effects were robust and persisted for at least two months. Although there was a slight performance decrease at the start of the second session, reasoners still provided more correct responses after two months than before the first training. Notably, in the pre-intervention block of the second session, performance for initial and final responses was higher in the video group than in the text group. This result suggests that the sustained training effect was stronger in the video than in the text group (mediated through a stronger initial training effect). Additionally, the second video intervention showed that reasoning performance could again be boosted.

Altogether, these findings are consistent with previous debiasing studies using text-based training (e.g., Boissin et al., 2021, 2022; Franiatte et al., 2024a, 2024b; Hoover & Healy, 2017) and highlight two keystone results. First, animated video training can lead to sound intuiting. Second, its effects are robust and persist for at least two months, with a (slightly) stronger training effect than text-based interventions. We believe that the present work can serve as a proof-of-principle for the video debiasing 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, if video training proves to be an effective method for debiasing, practical considerations should be taken into account when deciding between video and text training. That is, text-based materials are generally easier and faster to create, as they do not require designers or voice artists. Videos generally require more time and resources to produce. Additionally, to avoid hindering learning, the creation of these videos must adhere to certain recommendations from the multimedia and cognitive psychology literature. For example, it should consider human cognitive architecture and its constraints, avoid decorative animations that can increase extrinsic cognitive load, and meet user expectations and motivation (see Aalioui et al., 2022; Delmas, 2018). Nevertheless, videos can be particularly useful for specific audiences, such as teenagers, individuals with literacy challenges, or those with reading difficulties (Brown, 2007; Downs, 2014). Therefore, they could serve as a useful tool for debiasing the general public. Note that scholars interested in using our videos can access them freely on our OSF platform (https://osf.io/5fuh9).

Second, years of media comparison research have shown that the efficacy of animation depends not only on the medium used but also on the characteristics of the learners. That is, there are several “moderator factors” intrinsic to the learner that are known to play a key role in learning success. For instance, when watching videos, it is assumed that learning gains are stronger for learners with high spatial ability than for those with low spatial ability (Mayer, 2002). Similarly, learning gains may be better for learners with low prior knowledge than for those with high prior knowledge (Mayer, 2002). These individual differences could help explain variations in training success, such as those observed between improved and biased reasoners in our individual-level classification. Nevertheless, it may be worthwhile to examine in future work how individual differences could potentially account for variations in training accuracy. Against this backdrop, one could explore whether these variations are 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 video debiasing training.

Third, the current study focused on elementary logical principles in classic reasoning tasks (i.e., base-rate neglect, conjunction fallacy, and bat-and-ball tasks). These lab-based tasks are somewhat

artificial and context-specific (e.g., Janssen et al., 2021; Politzer et al., 2017; Prado et al., 2020). Arguably, people's erroneous personal beliefs in other contexts (e.g., climate change or medical contexts) might be more resistant to change. However, it has previously been showed that the effects of debiasing training hardly transfer across tasks or contexts (e.g., Boissin et al., 2021, 2022; Heijltjes et al., 2014; 2015; Van Peppen et al., 2021). Hence, future studies should ideally explore whether video debiasing effects extend to other reasoning tasks involving different logical principles and heuristics, or more ecological settings (e.g., Aalioui et al., 2022; Johan, 2024). However, it's important to note that mastering these elementary logical principles remain critical for sound reasoning in a wide range of situations. For instance, as the introductory example illustrates, base-rate neglect plays a significant role in the mistaken belief that airplane travel is unsafe (i.e., overlooking the extremely low probability of a crash given the high number of daily flights). Therefore, we believe it is essential to evaluate whether core logical principles can be effectively trained using classic reasoning tasks. At the same time, we acknowledge the need to further investigate the generalizability of these findings.

Finally, further research may build on these findings to refine interventions and optimize training methods. For instance, one well-known technique to boost learning outcomes is to have students retrieve the "to-be-learned" information from memory (e.g., Dunlosky et al., 2013; Fiorella & Mayer, 2016). Against this backdrop, Van Peppen et al. (2021) examined whether repeated retrieval practice improves critical thinking - particularly in reducing biased reasoning - and found a (non-significant) increase in average performance with more repetitions. Building upon this insight, one could attempt to further boost the present debiasing training efficacy by increasing the frequency of sessions or implementing retraining within shorter intervals (e.g., see Rawson & Dunlosky, 2022). Additionally, techniques such as adaptive learning—where training is tailored individually rather than applied uniformly—could also be considered (e.g., Adolphe et al., 2023; Van Gog et al., 2011). The optimal scheme remains to be explored here.

In conclusion, the present study suggests that animated video training is an effective tool to boost reasoning performance: It can help to improve accurate intuitive responding, its effects are robust, and tend to outperform mere text-based interventions. These findings serve as a proof-of-principle for the video debiasing training approach and warrants a wider and large-scale exploration and application of its potential.

Data availability statement

Raw data, analysis scripts, videos, and pre-registrations for these studies can be downloaded from our OSF page (https://osf.io/5fuh9).

Acknowledgements

We would like to thank Toscane Rabearisoa, Gabin Carrier, Leslie Stout, and Kaitlyn Cristelli for their precious help in designing the animated videos.

CRediT authorship contribution statement

Conceptualization: N. Franiatte; Data curation: N. Franiatte, Software: N. Franiatte; Investigation: N. Franiatte; Formal Analysis: N. Franiatte, E. Boissin; Methodology: N. Franiatte, E. Boissin, W. De Neys; Writing – Original Draft: N. Franiatte; Writing – Review and editing: N. Franiatte, E. Boissin, W. De Neys; Supervision: A. Delmas, W. De Neys; Funding Acquisition: A. Delmas, W. De Neys.

References

Adolphe, M., Pech, M., Sawayama, M., Maurel, D., Delmas, A., Oudeyer, P.-Y., & Sauzeon, H. (2023). *Exploring the Potential of Artificial Intelligence in Individualized Cognitive Training: a Systematic Review*.

Aalioui, L., Gouzi, F., & Tricot, A. (2022). Reducing cognitive load during video lectures in physiology with eye movement modeling and pauses: a randomized controlled study. *Advances in Physiology Education*, *46*(2), 288-296. https://doi.org/10.1152/advan.00185.2021

Andersson, L., Eriksson, J., Stillesjö, S., Juslin, P., Nyberg, L., & Wirebring, L.K. (2020). Neurocognitive processes underlying heuristic and normative probability judgments. *Cognition*, *196*. 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

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1-30. https://doi.org/10.1080/13546783.2018.1552194

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

Berney, S., & Bétrancourt, M. (2016). Does animation enhance learning? A meta-analysis. *Computers & Education*, *101*, 150-167. https://doi.org/10.1016/j.compedu.2016.06.005

Boissin, E., Caparos, S., & De Neys, W. (2023a). Examining the role of deliberation in de-bias training. *Thinking & Reasoning*, *30*(2), 327–355. https://doi.org/10.1080/13546783.2023.2259542

Boissin, E., Caparos, S., Hana, J.A., Bernard, C., & De Neys, W. (2024). Easy-fix attentional focus manipulation boosts the intuitive and deliberate use of base-rate information. *Memory & Cognition*, 1-13. https://doi.org/10.3758/s13421-024-01625-5

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

Brown, A., & Green, T. D. (2007). Video podcasting in perspective: The history, technology, aesthetics, and instructional uses of a new medium. *Journal of Educational Technology Systems*, *36*(1), 3-17. https://doi.org/10.2190/ET.36.1.b

Butcher, K.R. (2014). The multimedia principle. *The Cambridge handbook of multimedia learning*, *2*, 174-205.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations Still improve students' learning from text. *Educational Psychology Review*, *14*, 5–26. https://doi.org/10.1023/A:1013176309260

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

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

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

De Neys, W. (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

Delmas, A.A.D.M. (2018). *Conception et validation d'un jeu d'auto-apprentissage de connaissances sur l'asthme pour le jeune enfant: rôle de la motivation intrinsèque* (Doctoral dissertation, Université de Bordeaux).

Downs, J.S. (2014). Prescriptive scientific narratives for communicating usable science. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 13627–13633. https://doi.org/10.1073/pnas.1317502111

Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., & Willingham, D.T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58. https://doi.org/10.1177/1529100612453266

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

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

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*, 717-741. https://doi.org/10.1007/s10648-015-9348-9

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

Franiatte, N., Boissin, E., Delmas, A., & Neys, W. De. (2024b). Adieu Bias: Debiasing Intuitions Among French Speakers. *Psychologica Belgica*, *64*(1), 42–57. https://doi.org/10.5334/pb.1260

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

Frederick, S. (2005). Cognitive reflection and decision making. *The 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

Haferkamp, N., Kraemer, N.C., Linehan, C., & Schembri, M. (2011). Training disaster communication by means of serious games in virtual environments. *Entertainment Computing*, *2*(2), 81–88. https://doi.org/10.1016/j.entcom.2010.12.009

Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2014). Improving critical thinking: Effects of dispositions and instructions on economics students' reasoning skills. *Learning and Instruction*, *29*, 31-42. https://doi.org/10.1016/j.learninstruc.2013.07.003

Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2015). Unraveling the effects of critical thinking instructions, practice, and self-explanation on students' reasoning performance. *Instructional Science*, *43*, 487-506. https://doi.org/10.1007/s11251-015-9347-8

Höffler, T.N., & Leutner, D. (2007). Instructional animation versus static pictures: A meta-analysis. *Learning and Instruction*, *17*(6), 722–738. https://doi.org/10.1016/j.learninstruc.2007.09.013

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*. https://doi.org/10.1016/j.actpsy.2021.103322

Johan, J. (2024). *Developing and testing a checklist to improve scientific reasoning in complex decision tasks* (Doctoral dissertation, UNSW Sydney). https://doi.org/10.26190/unsworks/30672

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, & K. J. Holyoak (Eds.), The Cambridge handbook of thinking and reasoning (pp. 267–293). Cambridge University Press.

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

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

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

Mayer, R. E. (2002). Multimedia learning. In *Psychology of learning and motivation* (Vol. 41, pp. 85-139). Academic Press.

Mayer, R. E. (2005). *Cognitive theory of multimedia learning*. The Cambridge Handbook of Visuospatial Thinking/Cambridge University Press. https://doi.org/10.1017/cbo9780511816819.004

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. (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*, 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

R Core Team (2017) R: A Language and Environment for Statistical Computing. https://www.R-project.org/

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

Schnotz, W., & Rasch, T. (2005). Enabling, facilitating, and inhibiting effects of animations in multimedia learning: Why reduction of cognitive load can have negative results on learning. *Educational Technology Research and Development*, *53*(3), 47-58. https://doi.org/10.1007/BF02504797

Stanovich, K.E. (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), 645–726. 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

Tversky, B., Morrison, J.B., & Betrancourt, M. (2002). Animation: can it facilitate?. *International journal of human-computer studies*, *57*(4), 247-262. https://doi.org/10.1006/ijhc.2002.1017

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of concurrent monitoring on cognitive load and performance as a function of task complexity. *Applied Cognitive Psychology*, *25*(4), 584–587. https://doi.org/10.1002/acp.1726

Van Peppen, L., Verkoeijen, P., Heijltjes, A., Janssen, E., Koopmans, D., & Van Gog, T. (2018). Effects of self-explaining on learning and transfer of critical thinking skills. *Frontiers in education, 3*. https://doi.org/10.3389/feduc.2018.00100

Van Peppen, L.M., Verkoeijen, P.P., Heijltjes, A., Janssen, E., & Van Gog, T. (2021). Repeated retrieval practice to foster students' critical thinking skills. *Collabra: Psychology*, *7*(1), 28881. https://doi.org/10.1525/collabra.28881

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

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

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

Supplementary Material

A. Problems used

Items used in Studies 1-3 (i.e., base-rate, conjunction fallacy, and bat-and-ball tasks)

BB = bat-and-ball, BR = base-rate neglect, CF = conjunction fallacy
Pre = pre-intervention block, Int = intervention, Post = post-intervention-block

	Task & Block	Conflict version	No-conflict version
1	BB Pre	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 Pre	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 Pre	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 Pre	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 Pre	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 Pre	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 Pre	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 Pre	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 Int	A bat and ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?	
10	BB Int	A banana and an apple cost $1.40. The banana costs $1.00 more than the apple. How much does the apple cost?	
11	BB Int	A magazine and a banana cost $2.60 in total. The magazine costs $2.00 more than the banana. How much does the banana cost?	
12	BB Post	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?

13	BB Post	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?
14	BB Post	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?
15	BB Post	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?
16	BB Post	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?
17	BB Post	For a sports event, organizers have invited 530 players and coaches. There are 500 more players than coaches. How many coaches are there?	For a sports event, organizers have invited 510 players and coaches. There are 500 players. How many coaches are there in this event?
18	BB Post	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?
19	BB Post	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?
1	BR Pre	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?
2	BR Pre	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?
3	BR Pre	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?
4	BR Pre	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?

5	BR Pre	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?
6	BR Pre	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?
7	BR Pre	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?
8	BR Pre	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?
9	BR Int	This study contains lab technicians and politicians. Person 'F' is dishonest. There are 996 lab technicians and 4 politicians. *Is Person 'F' more likely to be:* - A lab technician? - A politician?	
10	BR Int	This study contains Hollywood celebrities and bakers. Person 'C' is rich. There are 5 Hollywood celebrities and 995 bakers. *Is Person 'C' more likely to be:* - A Hollywood celebrity? - A baker?	
11	BR Int	This study contains boxers and kindergarten teachers. Person 'V' is kind. There are 995 boxers and 5 kindergarten teachers. *Is Person 'V' more likely to be:* - A boxer? - A kindergarten teacher?	
12	BR Post	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?
13	BR	This study contains accountants and boys. Person 'H' is immature.	This study contains accountants and boys. Person 'H' is immature.

	Post	There are 997 accountants and 3 boys. *Is Person 'H' more likely to be:* - An accountant? - A boy?	There are 3 accountants and 997 boys. *Is Person 'H' more likely to be:* - An accountant? - A boy?
14	BR Post	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?
15	BR Post	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?
16	BR Post	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?
17	BR Post	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?
18	BR Post	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?
19	BR Post	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?	This study contains secretaries and telemarketers. Person 'J' is persuasive. There are 5 secretaries and 995 telemarketers. *Is Person 'J' more likely to be:* - A secretary? - A telemarketer?
1	CF Pre	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

2	CF Pre	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
3	CF Pre	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
4	CF Pre	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
5	CF Pre	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
6	CF Pre	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
7	CF Pre	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
8	CF Pre	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
9	CF Int	Tracy, 45, has previously studied synchronized swimming and likes the beach. Is it most probable that the described person is:	

		- A plumber - A celebrity DJ - A plumber and a tanner - A plumber and a diver	
10	CF Int	Sloan, 39, has previously studied masonry and likes tattoos. Is it most probable that the described person is: - A nanny - A deputy - A nanny and a cat lover - A nanny and a hard rock lover	
11	CF Int	Henri, 36, has previously studied journalism and likes gossip. Is it most probable that the described person is: - A forest ranger - A mine-clearer - A forest ranger and a handyman - A forest ranger and a tabloid reader	
12	CF Post	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
13	CF Post	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
14	CF Post	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
15	CF Post	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
16	CF Post	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

17	CF Post	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
18	CF Post	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 - A house painter and an iPad owner - A house painter	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 and a carpet weaver
19	CF Post	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 Study 4-6 (i.e., base-rate re-test, conjunction fallacy re-test, and bat-and-ball re-test):

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

Pre = pre-intervention block, Int = intervention, Post = post-intervention-block

	Task & Block	**Conflict version**	**No-conflict version**
1	BB Pre	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 Pre	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 Pre	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 Pre	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 Pre	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 Pre	The kitchen in a restaurant has 240 plates and pans in total. There are 200 more plates than pans.	The kitchen in a restaurant has 250 plates and pans in total. There are 200 plates.

		How many pans are there?	How many pans are there?
7	BB Pre	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 Pre	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 Int	A bat and ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?	
10	BB Int	A banana and an apple cost $1.40. The banana costs $1.00 more than the apple. How much does the apple cost?	
11	BB Int	A magazine and a banana cost $2.60 in total. The magazine costs $2.00 more than the banana. How much does the banana cost?	
12	BB Post	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?
13	BB Post	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?
14	BB Post	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?
15	BB Post	In a city park there are 390 skateboarders and pedestrians. There are 300 more skateboarders than pedestrians. How many pedestrians are there?	In a city park there are 640 skateboarders and pedestrians. There are 600 skateboarders. How many pedestrians are there in this park?
16	BB Post	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?
17	BB Post	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?
18	BB Post	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?
19	BB Post	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?
1	BR Pre	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?
2	BR Pre	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?
3	BR Pre	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?
4	BR Pre	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?
5	BR Pre	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?
6	BR Pre	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?
7	BR Pre	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?
8	BR Pre	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?
9	BR Int	This study contains lab technicians and politicians. Person 'F' is dishonest. There are 996 lab technicians and 4 politicians. *Is Person 'F' more likely to be:* - A lab technician? - A politician?	

10	BR Int	This study contains Hollywood celebrities and bakers. Person 'C' is rich. There are 5 Hollywood celebrities and 995 bakers. *Is Person 'C' more likely to be:* - A Hollywood celebrity? - A baker?	
11	BR Int	This study contains boxers and kindergarten teachers. Person 'V' is kind. There are 995 boxers and 5 kindergarten teachers. *Is Person 'V' more likely to be:* - A boxer? - A kindergarten teacher?	
12	BR Post	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?
13	BR Post	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?
14	BR Post	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?
15	BR Post	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?
16	BR Post	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?
17	BR Post	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?	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 lawyer?	- A gardener?
18	BR Post	This study contains women and drummers. Person 'M' is sensitive. There 4 women and 996 drummers. Is Person 'M' more likely to be: - A drummer? - A woman?	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?
19	BR Post	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?
1	CF Pre	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
2	CF Pre	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
3	CF Pre	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
4	CF Pre	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
5	CF Pre	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

6	CF Pre	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
7	CF Pre	Jenny, 33, 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	Grady, 37, has previously studied political science and likes local politics. 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
8	CF Pre	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
9	CF Int	Tracy, 45, has previously studied synchronized swimming and likes the beach. Is it most probable that the described person is: - A plumber - A celebrity DJ - A plumber and a tanner - A plumber and a diver	
10	CF Int	Sloan, 39, has previously studied masonry and likes tattoos. Is it most probable that the described person is: - A nanny - A deputy - A nanny and a cat lover - A nanny and a hard rock lover	
11	CF Int	Henri, 36, has previously studied journalism and likes gossip. Is it most probable that the described person is: - A forest ranger - A mine-clearer - A forest ranger and a handyman - A forest ranger and a tabloid reader	
12	CF Post	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
13	CF Post	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	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 stock speculator - A rock star - A bus driver	- A bus driver and a social democrat - A social democrat - A rock star
14	CF Post	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
15	CF Post	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
16	CF Post	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
17	CF Post	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
18	CF Post	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
19	CF Post	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

Text explanations given during the intervention block

For the base-rate task (Study 1 and Study 4)

Question 1:

“This study contains lab technicians and politicians. Person 'F' is dishonest. There are 996 lab technicians and 4 politicians. Is Person 'F' more likely to be a lab technician or a politician?”

Text explanation 1:
“The correct answer to the previous problem is that person F is most likely a “Lab technician”. Most people think the answer is a “Politician”, but this answer is wrong.

Most people base their answer solely on the description (“Person F is dishonest”). If this were the only information given, this answer would be correct, as it is likely that there are more dishonest politicians in the world than dishonest lab technicians.

However, in the problem, you also got information about the specific number of lab technicians and politicians in the group from which person F got drawn. You were informed that person F was drawn randomly from a group with 996 lab technicians and only 4 politicians. Since there are so many more lab technicians in the group than politicians (almost 200 times more!), it becomes more likely that person F is a lab technician. After all, although politicians might generally be more dishonest than lab technicians, some lab technicians are dishonest.

If you combine this with the vastly larger number of lab technicians in the group, it will be more plausible that you’re dealing with a dishonest lab technician.”

Question 2:
“This study contains Hollywood celebrities and bakers. Person 'C' is rich. There are 5 Hollywood celebrities and 995 bakers. Is Person 'C' more likely to be a Hollywood celebrity or a baker?”

Text explanation 2:
“The correct answer to the previous problem is that person C is most likely a baker. Most people think the answer is a “Hollywood celebrity”, but this answer is wrong.

Most people base their answer solely on the description (“Person C is rich”). If this were the only information given, this answer would be correct, as it is likely that there are more rich Hollywood celebrities in the world than rich bakers.

However, in the problem, you also got information about the specific number of bakers and Hollywood celebrities in the group from which person C got drawn. You were informed that person C was drawn randomly from a group with 995 bakers and only 5 Hollywood celebrities. Since there are so many more bakers in the group than Hollywood celebrities (almost 200 times more!), it becomes more likely

that person C is a baker. After all, while Hollywood celebrities are generally wealthier than bakers, some bakers are rich.

If you combine this with the vastly larger number of bakers in the group, it will be more plausible that you're dealing with a rich baker."

Question 3:

"This study contains boxers and kindergarten teachers. Person 'V' is kind. There are 995 boxers and 5 kindergarten teachers. Is Person 'V' more likely to be a boxer or a kindergarten teacher?"

Text explanation 3:

"The correct answer to the previous problem is that person V is most likely a baker. Most people think the answer is a "Kindergarten teacher", but this answer is wrong.

Most people base their answer solely on the description ("Person V is kind"). If this were the only information given, this answer would be correct, as it is likely that there are more kind kindergarten teachers in the world than kind boxers.

However, in the problem, you also got information about the specific number of boxers and kindergarten teachers in the group from which person V got drawn. You were informed that person V was drawn randomly from a group with 995 boxers and only 5 kindergarten teachers. Since there are so many more boxers in the group than kindergarten teachers (almost 200 times more!), it becomes more likely that person V is a boxer. After all, although kindergarten teachers might in general be kinder than boxers, some boxers are kind.

If you combine this with the vastly larger number of boxers in the group, it will be more plausible that you're dealing with a kind boxer."

For the conjunction fallacy task (Study 2 and Study 5)

Question 1:

"Tracy, 45, has previously studied synchronized swimming and likes the beach. Is it most probable that the described person is a plumber, a celebrity DJ, a plumber and a tanner, or a plumber and a diver?"

Text explanation 1:

"The correct answer to the previous problem is that Tracy is most likely "a plumber". Most people think that the answer is "a plumber and a diver" but this answer 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 Tracy's educational background and interests, it seems more realistic to think of Tracy as a plumber and a diver rather than only a plumber. Simply because adding that Tracy is also a diver is more in line with our representation of someone who has studied synchronised swimming and likes the beach, rather than Tracy only being a plumber.

If one of the proposed answers had been a diver, then this reasoning would probably be correct. However, in this problem, the option "a diver" is presented together with another event: "a plumber". Now the statistical probability that Tracy is a plumber is higher than the probability that Tracy is a plumber **AND** a diver. 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 plumber". Some plumbers will also be divers, others will not be divers. The group of people who are plumbers and divers is a subgroup of the group of all plumbers. Hence, there will always be more people who are simply plumbers than people who are plumbers and in addition also divers."

Question 2:

"Sloan, 39, has previously studied masonry and likes tattoos. Is it most probable that the described person is a nanny, a deputy, a nanny and a cat lover, or a nanny and a hard rock lover?"

Text explanation 2:

"The correct answer to the previous problem is that Sloan is most likely "a nanny". Most people think that the answer is "a nanny and a hard rock lover" but this answer 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 Sloan's educational background and interests, it seems more realistic to think of Sloan as a nanny and a hard rock lover rather than only a nanny. Simply because adding that Sloan is also a hard rock lover is more in line with our representation of someone who has studied masonry and likes tattoos, rather than Sloan only being a nanny.

If one of the proposed answers had been "a hard rock lover" then this reasoning would probably be correct. However, in this problem, the option "a hard rock lover" is presented together with another event: "a nanny". Now the statistical probability that Sloan is a nanny is higher than the probability that Sloan is a nanny **AND** a hard rock lover. 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 nanny". Some nannies will also be hard rock lovers, others will not be hard rock lovers. The group of people who are nannies and hard rock lovers is a subgroup of the group of all nannies. Hence, there will always be more people who are simply nannies than people who are nannies and in addition also hard rock lovers."

Question 3:

"Henri, 36, has previously studied journalism and likes gossip. Is it most probable that the described person is a forest ranger, a mine-clearer, a forest ranger and a handyman, or a forest ranger and a tabloid reader?"

Text explanation 3:

"The correct answer to the previous problem is that Henri is most likely "a forest ranger". Most people think that the answer is "a forest ranger and a tabloid reader" but this answer 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 Henri's educational background and interests, it seems more realistic to think of Henri as a forest ranger and a tabloid reader rather than only a forest ranger. Simply because adding that Henri is also a tabloid reader is more in line with our representation of someone who has studied journalism and likes gossip, rather than Henri only being a forest ranger.

If one of the proposed answers had been "a tabloid reader" then this reasoning would probably be correct. However, in this problem the option "a tabloid reader" is presented together with another event: "a forest ranger". Now the statistical probability that Henri is a forest ranger is higher than the probability that Henri is a forest ranger **AND** a tabloid reader. 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 forest ranger". Some forest rangers will also be tabloid readers, others won't be. The group of people who are forest rangers and tabloid readers is a subgroup of the group of all forest rangers. Hence, there will always be more people who are simply forest rangers than people who are forest rangers and in addition also tabloid readers."

For the bat-and-ball task (Study 3 and Study 6)

Question 1:

"A bat and ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?"

Text explanation 1:

"The correct answer to the previous problem is 5 cents. Most people think it is 10 cents, but this answer is wrong.

If the ball costs 10 cents the bat would cost $1.10 (as it costs $1.00 more than the ball); both together, they would then cost $1.20.

However, the problem said they cost $1.10 together.

The correct response is that the ball costs 5 cents, the bat $1.05 so together they cost $1.10 ($0.05 + $1.05 = $1.10)."

Question 2:

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

Text explanation 2:

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

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

However, the problem said they cost $1.40 together.

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

Question 3:

"A magazine and a banana cost $2.60 in total. The magazine costs $2.00 more than the banana. How much does the banana cost?"

Text explanation 3:

"The correct response is 30 cents. Most people are tempted to answer 60 cents, but this is wrong.

If the banana costs 60 cents, the magazine would cost $2.60 (as it costs two dollars more than the banana); both together, they would then cost $3.20.

However, the problem said they cost $2.60 together.

The correct answer is that the banana costs 30 cents, the magazine $2.30 so together they cost $2.60 ($0.30 + $2.30 = $2.60).”

B. Conjunction fallacy problems: Frequency of each individual response option in Study 2 (Figure S1) and Study 5 (Figure S2)

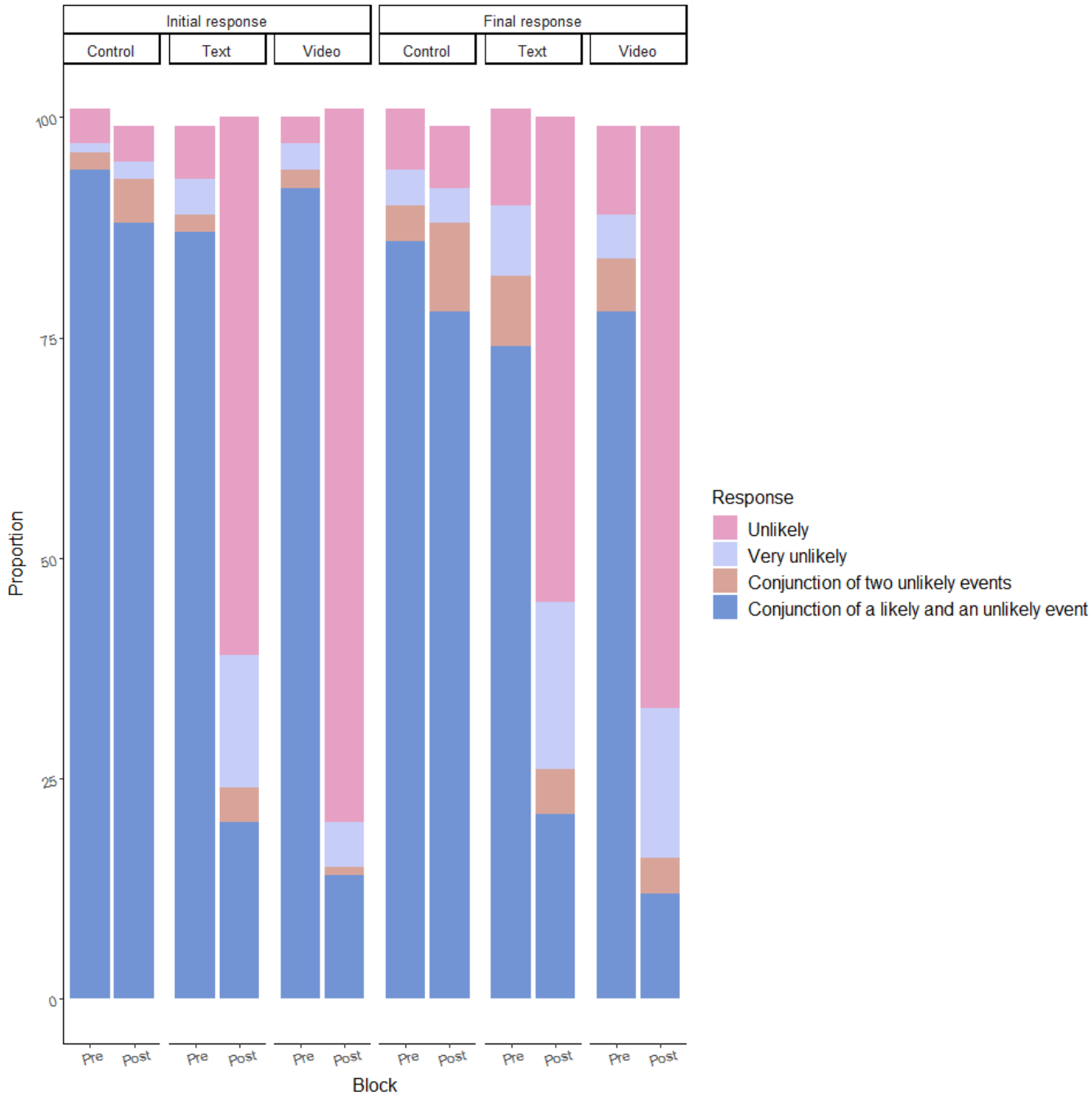

Figure S1. Frequency of each individual response option in Study 2 (conjunction fallacy, first training session) for the initial and the final conflict responses, before and after the intervention in the control, text, and video groups.

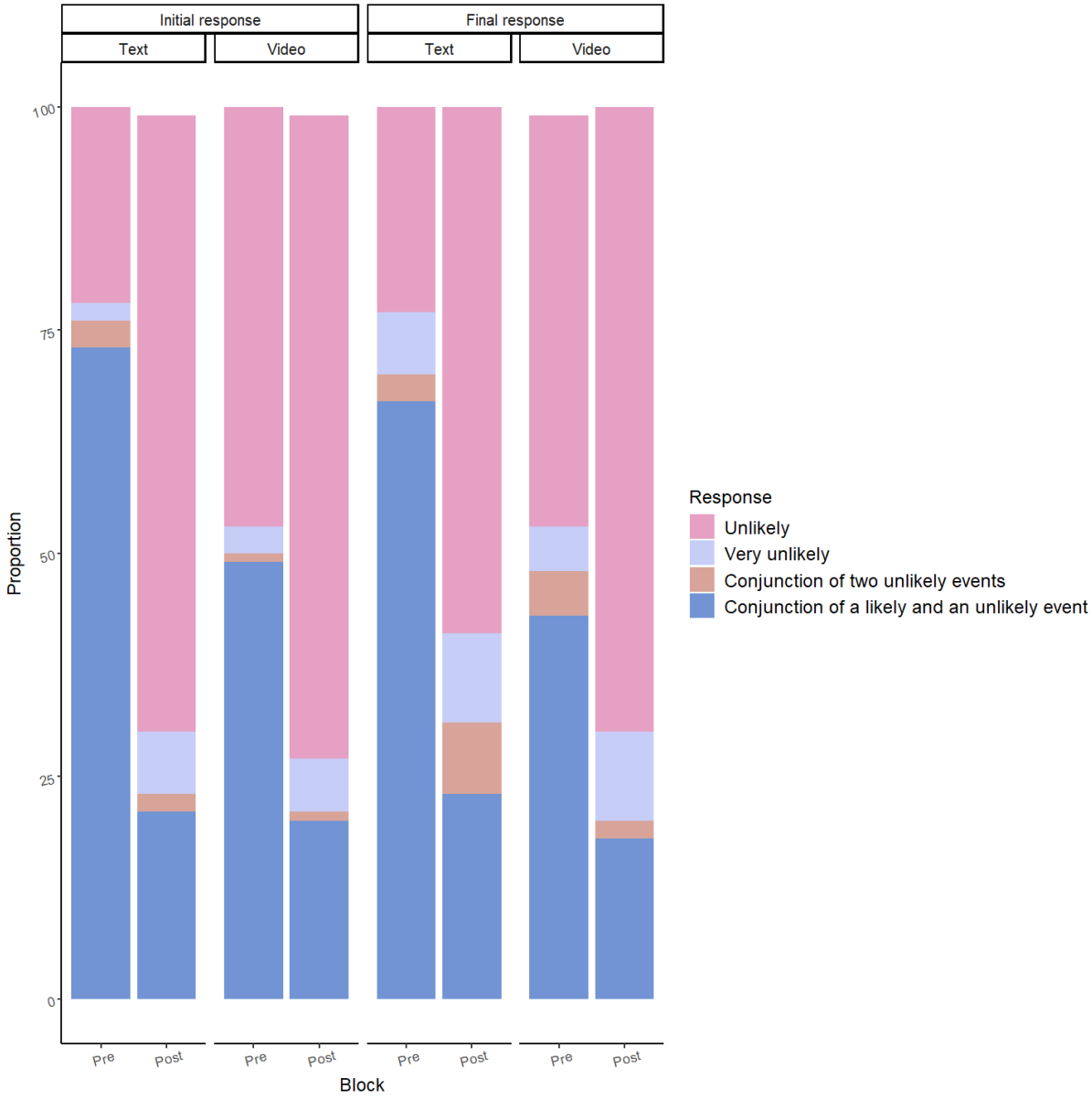

Figure S2. Frequency of each individual response option in Study 5 (conjunction fallacy, re-test training session) for the initial and the final conflict responses, before and after the intervention in the text and video groups.

C. Justifications

Table S1. Frequency of different types of justifications for the final base-rate (BR), conjunction fallacy (CF), and bat-and-ball (BB) conflict problems and all tasks combined (All) for the control, text, and video groups during the post-intervention of the first training session (Studies 1-3).

Task	Justification	Control group		Text group		Video group	
		Correct response (n = 55)	*Incorrect response (n = 94)*	*Correct response (n = 116)*	*Incorrect response (n = 34)*	*Correct response (n = 126)*	*Incorrect response (n = 24)*
All	Math - Correct	42	2	71	1	88	/
	Math – Incorrect or Unspecified	/	37	5	15	10	8
	Guess	2	3	12	3	3	2
	Intuitions	7	38	23	12	12	8
	Other	4	14	5	3	13	6
BR	Math - Correct	23	1	34	1	39	/
	Math – Incorrect or Unspecified	/	2	1	2	/	/
	Guess	1	/	3	/	1	/
	Intuitions	2	13	5	2	2	1
	Other	3	5	2	/	6	1
CF	Math - Correct	1	/	10	/	19	/
	Math – Incorrect or Unspecified	/	17	3	3	9	3
	Guess	/	1	8	1	2	/
	Intuitions	3	20	17	4	7	2
	Other	1	7	2	2	5	2
BB	Math - Correct	18	1	27	/	30	/
	Math – Incorrect or Unspecified	/	18	1	10	1	5
	Guess	1	2	1	2	/	2
	Intuitions	2	5	1	6	3	5
	Other	/	2	1	1	2	3

Table S2. Frequency of different types of justifications for the final base-rate (BR), conjunction fallacy (CF), and bat-and-ball (BB) conflict problems and all tasks combined (All) for the text and video groups during the post-intervention of the re-test training session (Studies 4-6).

Task	**Justification**	**Text group**		**Video group**	
		Correct response (n = 97)	*Incorrect response (n = 18)*	*Correct response (n = 89)*	*Incorrect response (n = 21)*
All Re-test	Math - Correct	58	/	55	/
	Math – Incorrect or Unspecified	9	12	13	10
	Guess	4	1	5	1
	Intuitions	16	3	7	9
	Other	10	2	9	1
BR Re-test	Math - Correct	26	/	26	/
	Math – Incorrect or Unspecified	2	/	3	/
	Guess	1	/	2	/
	Intuitions	5	/	2	/
	Other	3	/	3	/
CF Re-test	Math - Correct	9	/	11	/
	Math – Incorrect or Unspecified	4	6	4	4
	Guess	2	1	3	/
	Intuitions	9	2	5	5
	Other	6	1	3	/
BB Re-test	Math - Correct	23	/	18	/
	Math – Incorrect or Unspecified	3	6	6	6
	Guess	1	/	/	1
	Intuitions	2	1	/	4
	Other	1	1	3	1

D. Accuracy for no-conflict problems

Table S3. Average accuracy (%) for the no-conflict problems (SD) for each task (BR, CF, BB) and combined (All) in the first training session (Studies 1-3). BR = base-rate neglect, CF = conjunction fallacy tasks, BB = bat-and-ball, All = the composite mean across the three tasks.

Task	Group	Initial response		Final response	
		Pre-intervention	Post-intervention	Pre-intervention	Post-intervention
All	Control	86.8 (23.3)	86.0 (26.7)	87.7 (24.8)	90.2 (22.5)
	Text	88.7 (22.3)	93.9 (15.8)	89.8 (20.8)	97.0 (11.1)
	Video	88.0 (22.7)	90.4 (23.1)	88.4 (22.5)	94.6 (19.5)
BR	Control	94.2 (17.4)	89.1 (24.8)	96.1 (15.7)	94.7 (14.0)
	Text	98.7 (6.6)	97.1 (8.8)	98.7 (6.6)	100.0 (0)
	Video	91.8 (19.7)	95.3 (20.4)	95.8 (13.5)	98.0 (14.3)
CF	Control	73.5 (27.9)	75.7 (30.6)	72.2 (29.6)	81.2 (28.2)
	Text	70.1 (28.5)	89.2 (21.9)	74.0 (27.3)	93.1 (16.2)
	Video	75.7 (29.2)	92.9 (15.9)	72.5 (29.2)	96.3 (15.6)
BB	Control	92.7 (17.2)	93.4 (20.9)	94.9 (19.1)	94.9 (20.4)
	Text	97.7 (10.5)	95.6 (12.5)	97.0 (11.4)	97.8 (9.1)
	Video	96.1 (10.5)	83.3 (29.3)	96.4 (11.6)	89.7 (25.6)

Table S4. Average accuracy (%) for the no-conflict problems (SD) for each task (BR, CF, BB) and combined (All) in the re-test training session (Studies 4-6). BR = base-rate neglect, CF = conjunction fallacy tasks, BB = bat-and-ball, All = the composite mean across the three tasks.

Task	Group	Initial response		Final response	
		Pre-intervention	Post-intervention	Pre-intervention	Post-intervention
All Re-test	Text	78.2 (29.4)	85.3 (25.9)	81.5 (27.6)	85.5 (26.6)
	Video	82.4 (27.0)	85.5 (26.3)	85.6 (26.1)	89.6 (23.8)
BR Re-test	Text	93.5 (13.2)	83.3 (19.4)	94.6 (12.9)	85.1 (21.9)
	Video	94.9 (11.7)	80.8 (24.1)	99.3 (4.2)	86.6 (20.1)
CF Re-test	Text	62.7 (36.9)	87.8 (27.1)	68.3 (36.3)	83.8 (27.7)
	Video	76.7 (33.2)	86.2 (27.1)	77.9 (32.9)	90.2 (26.0)
BB Re-test	Text	79.6 (23.7)	84.5 (30.3)	82.7 (20.9)	87.6 (30.0)
	Video	76.1 (27.5)	89.1 (27.4)	79.9 (26.5)	91.7 (25.2)

E. Individual level direction of change in Studies 1-6

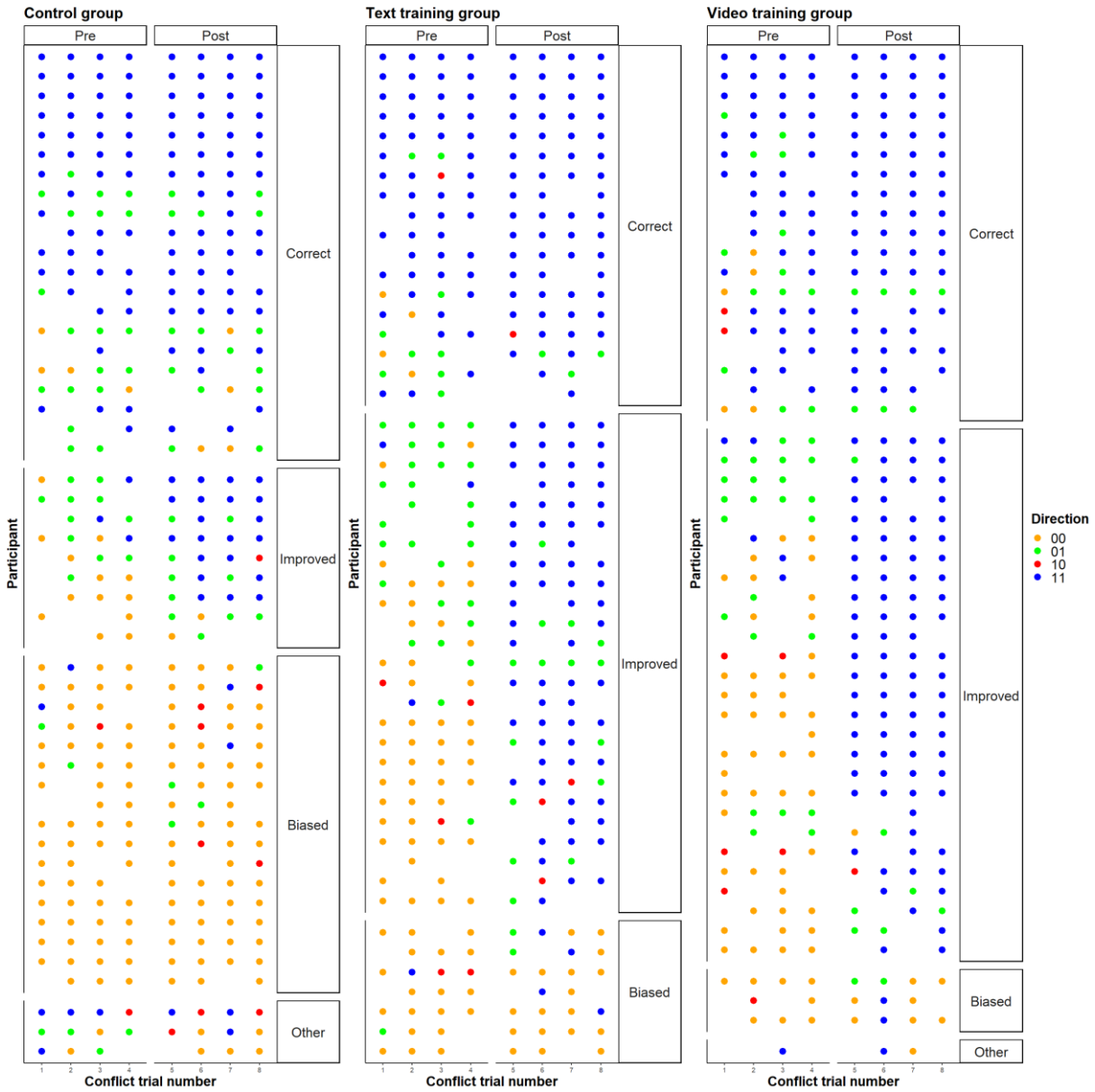

Figure S3. Individual level direction of change (each row represents one participant) and classification in Study 1 (base-rate neglect task). Due to the exclusion of missed deadline and load trials (see 2.1.7 Trial Exclusion), not all participants contributed 8 analysable trials.

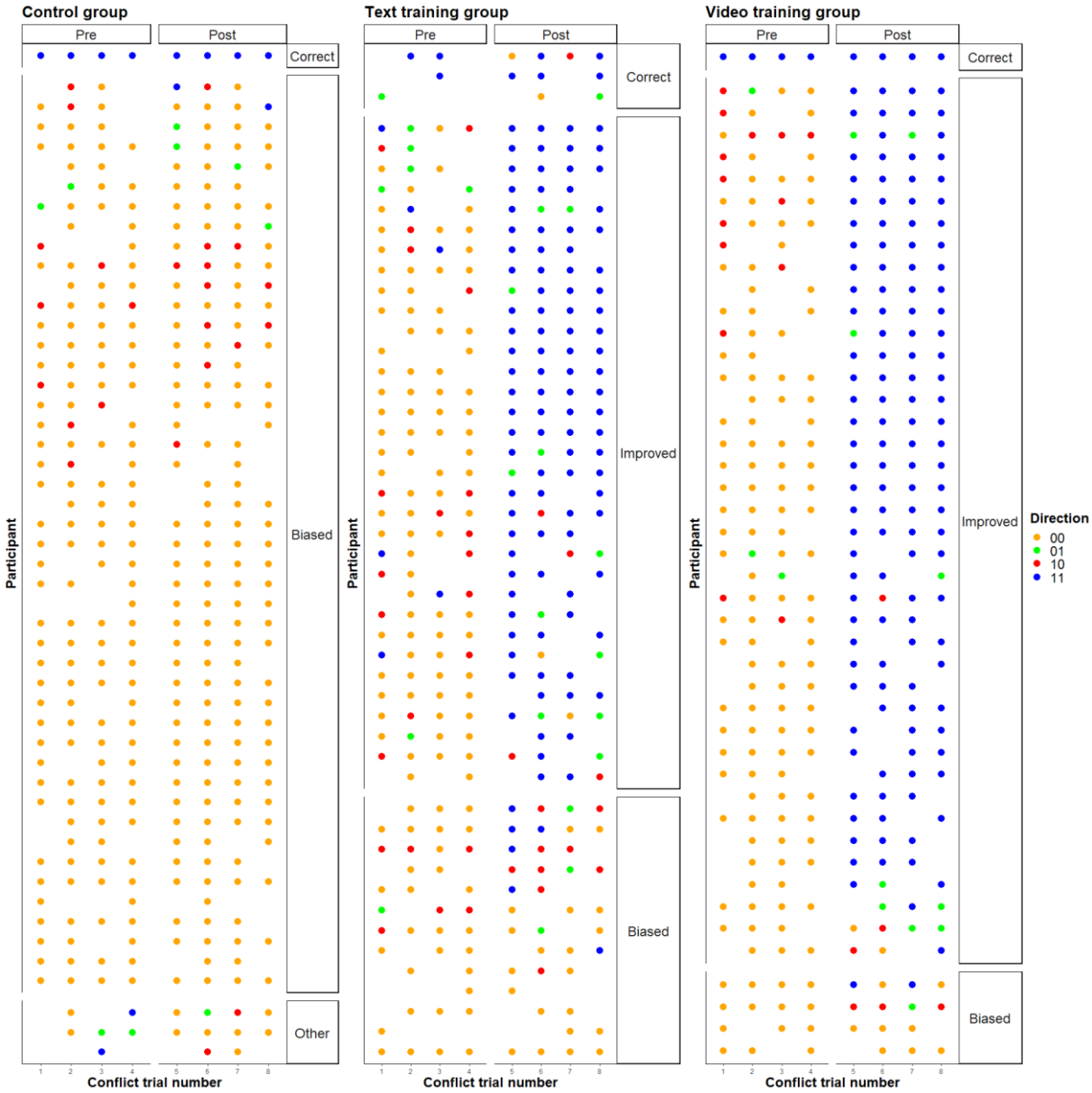

Figure S4. Individual level direction of change (each row represents one participant) and classification in Study 2 (conjunction fallacy task). Due to the exclusion of missed deadline and load trials (see 2.1.7 Trial Exclusion), not all participants contributed 8 analysable trials.

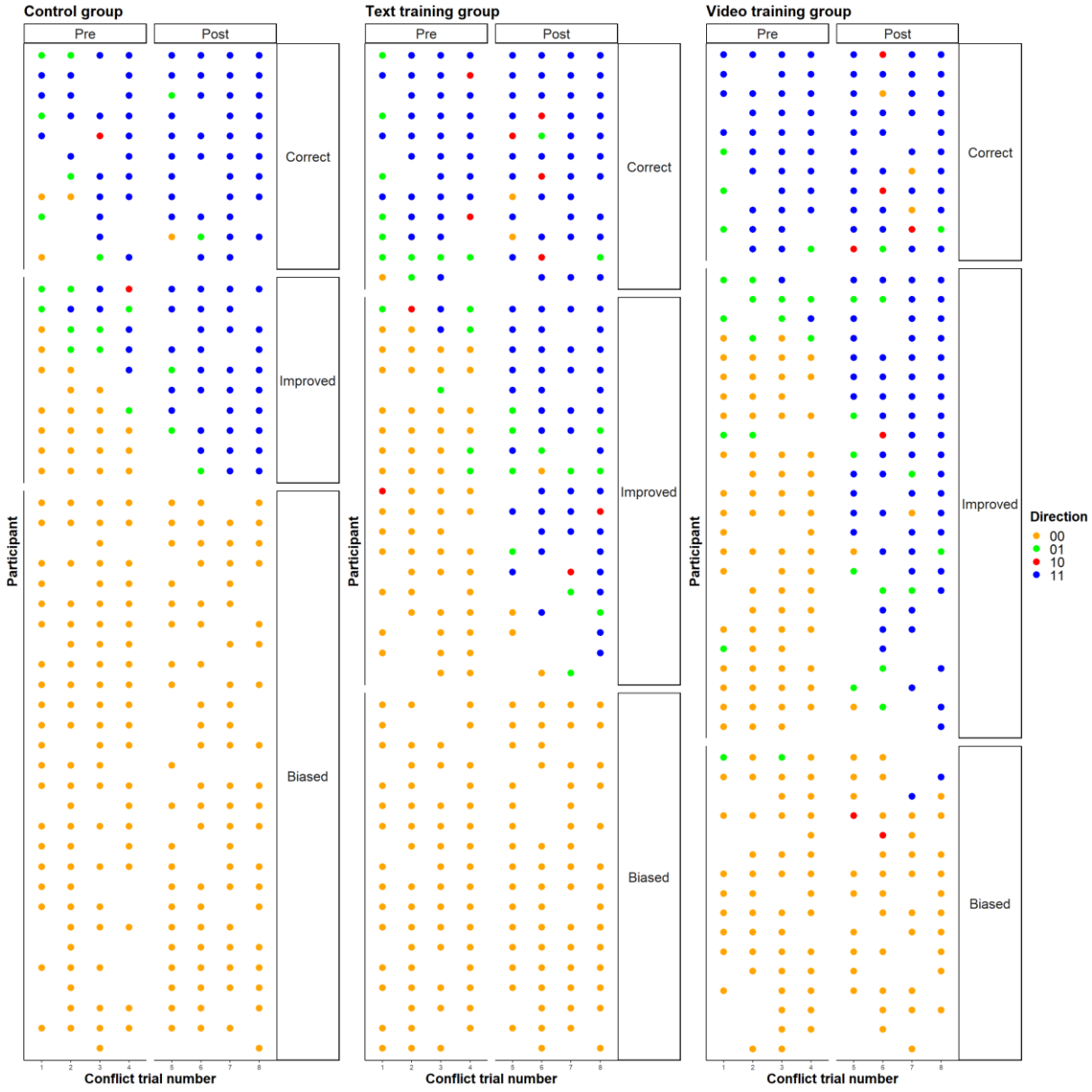

Figure S5. Individual level direction of change (each row represents one participant) and classification in Study 3 (bat-and-ball task). Due to the exclusion of missed deadline and load trials (see 2.1.7 Trial Exclusion), not all participants contributed 8 analysable trials.

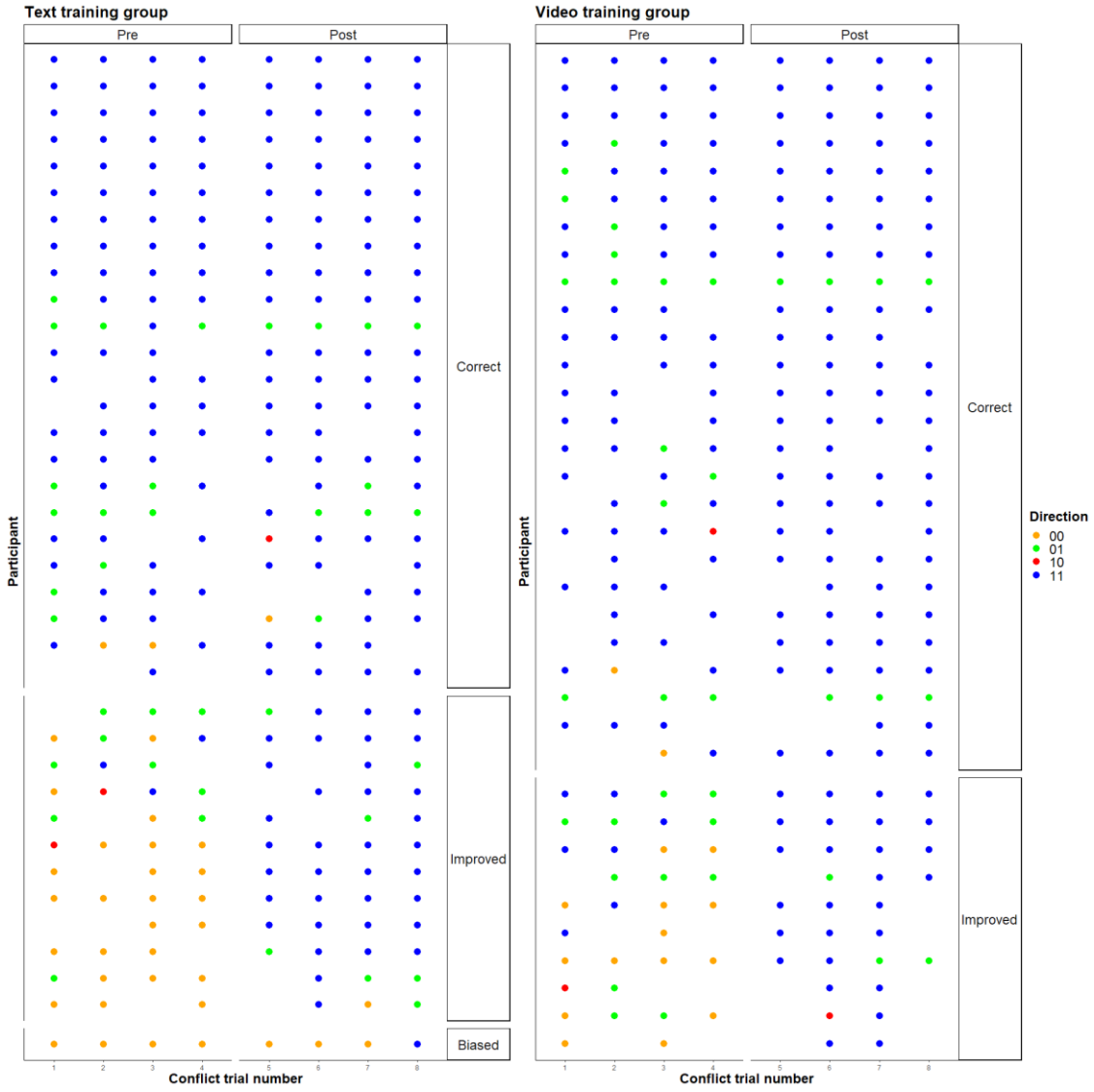

Figure S6. Individual level direction of change (each row represents one participant) and classification in Study 4 (base-rate re-test). Due to the exclusion of missed deadline and load trials (see 2.1.7 Trial Exclusion), not all participants contributed 8 analysable trials.

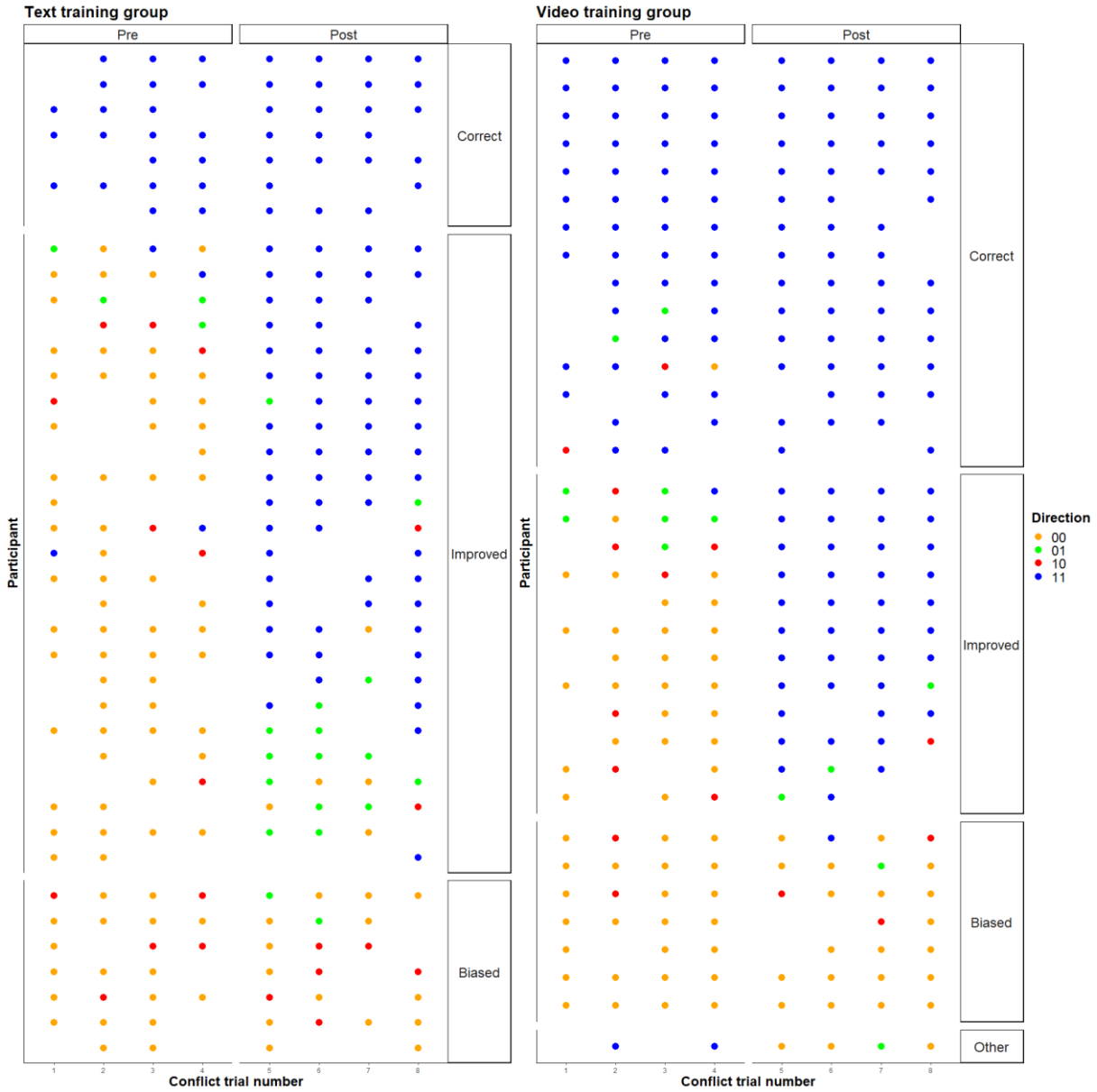

Figure S7. Individual level direction of change (each row represents one participant) and classification in Study 5 (conjunction fallacy re-test). Due to the exclusion of missed deadline and load trials (see 2.1.7 Trial Exclusion), not all participants contributed 8 analysable trials.

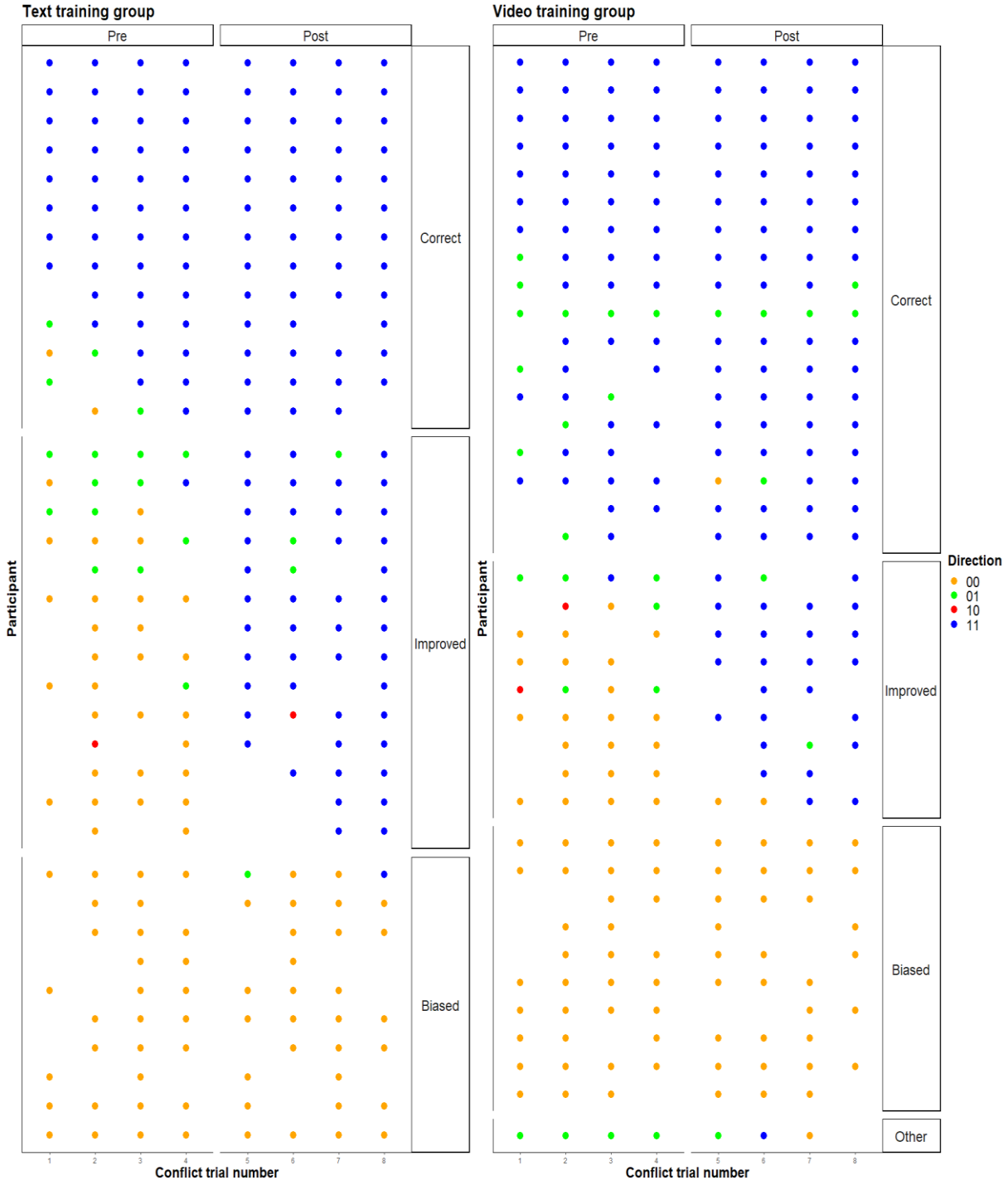

Figure S8. Individual level direction of change (each row represents one participant) and classification in Study 6 (bat-and-ball re-test). Due to the exclusion of missed deadline and load trials (see 2.1.7 Trial Exclusion), not all participants contributed 8 analysable trials.

F. Conflict detection analyses

Previous debiasing studies (Boissin et al., 2021, 2022; Franiatte et al., 2024a), reported a trend towards an improved conflict detection index following the text-based training, specifically for the bat-and-ball and base-rate tasks. This trend was not observed in the conjunction fallacy task. Additionally, some authors argued against the use of the index with the specific conjunction format we adopted (see Aczel et al., 2016; Scherer et al., 2017). Therefore, we decided to analyse the results for each task separately.

Table S5. Conflict detection results in Studies 1-3. Percentage of mean difference in confidence ratings (SD) between correct no-conflict and incorrect conflict problems on each reasoning task: Base-rate neglect (BR), conjunction fallacy (CF), and bat-and-ball (BB).

Task	Group	Initial response	
		Pre-intervention	Post-intervention
BR	Video	3.6 (8.9)	8.8 (11.3)
	Text	8.9 (15.1)	16.0 (21.8)
	Control	8.9 (19.3)	8.6 (16.5)
CF	Video	3.1 (16.9)	13.6 (23.2)
	Text	10.6 (19.5)	13.3 (24.6)
	Control	4.0 (1801)	2.4 (18.0)
BB	Video	14.4 (23.4)	18.6 (32.2)
	Text	6.0 (20.8)	5.8 (15.4)
	Control	0.0 (14.0)	5.4 (18.1)

Table S6. Conflict detection results in Studies 4-6. Percentage of mean difference in confidence ratings (SD) between correct no-conflict and incorrect conflict problems on each reasoning task: Base-rate neglect (BR), conjunction fallacy (CF), and bat-and-ball (BB).

Task	Group	Initial response	
		Pre-intervention	Post-intervention
BR Re-test	Video	6.75 (8.81)	48.0 (43.2)
	Text	6.7 (18.9)	21.2 (36.4)
CF Re-test	Video	-4.1 (28.9)	11.4 (35.2)
	Text	3.5 (19.4)	6.3 (20.3)
BB Re-test	Video	21.7 (33.0)	17.4 (30.8)
	Text	10.0 (24.0)	19.0 (23.8)

G. Predictive conflict detection analyses

Since previous works reported trends towards a predictive conflict detection effect for bat-and-ball and base-rate tasks, but not for the conjunction fallacy task (see Boissin et al., 2022; Franiatte et al., 2024a), we decided to analyse the results for each task independently.

Table S7. Predictive Conflict Detection results in Studies 1-3. 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 video and text groups, and for each reasoning task: Base-rate neglect (BR), conjunction fallacy (CF), and bat-and-ball (BB).

Task	Group	Label (N)	Initial response – Session 1
			Pre-intervention
BR	Video	Improved (n = 27)	16.1 (29.0)
		Biased (n = 3)	-0.7 (12.0)
	Text	Improved (n = 25)	17.5 (22.2)
		Biased (n = 7)	2.1 (2.9)
CF	Video	Improved (n = 37)	10.1 (18.0)
		Biased (n = 4)	1.3 (20.8)
	Text	Improved (n = 33)	8.8 (19.5)
		Biased (n = 12)	-1.2 (14.5)
BB	Video	Improved (n = 24)	17.5 (26.0)
		Biased (n = 16)	7.4 (17.3)
	Text	Improved (n = 19)	8.2 (21.2)
		Biased (n = 18)	2.4 (10.3)

Table S8. Predictive Conflict Detection results in Studies 4-6. 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 video and text groups, and for each reasoning task: Base-rate neglect (BR), conjunction fallacy (CF), and bat-and-ball (BB).

Task	Group	Label (N)	Initial response – Session 1
			Pre-intervention
BR Re-test	Video	Improved (n = 10)	23.8 (38.0)
		Biased (n = 0)	/
	Text	Improved (n = 12)	0.7 (15.3)
		Biased (n = 1)	0.0 (0.0)
CF Re-test	Video	Improved (n = 12)	-11.3 (21.4)
		Biased (n = 4)	N 9.5 (17.3)
	Text	Improved (n = 22)	4.4 (19.4)
		Biased (n = 6)	2.7 (15.8)
BB Re-test	Video	Improved (n = 6)	34.3 (35.6)
		Biased (n = 10)	4.1 (8.9)
	Text	Improved (n = 14)	11.0 (15.5)
		Biased (n = 10)	7.8 (26.5)

H. Ratings

Table S9. Ratings of training interventions by participants of the video and text conditions. As a reminder, note that at the end of each task, participants in the video and text groups were asked to rate on a scale from 0 (not at all) to 10 (extremely) the clarity, enjoyment, and informativeness of the explanations they received.

Question	Group	Mean (SD)
How clear did you find the explanations?	Video	8.9 (1.7)
	Text	8.5 (2.0)
To what extent did you enjoy the explanations?	Video	7.0 (2.7)
	Text	7.3 (2.5)
To what extent did you find the explanations informative (you felt you learned something)?	Video	6.6 (3.2)
	Text	7.0 (2.8)

NB: Only the clarity ratings significantly distinguished the video from the text-based training, with a significant difference, $t(287) = 2.05$, $p = .04$. The other measures, enjoyment, $t(290) = 0.83$, $p = .41$, and informativeness, $t(287) = 1.24$, $p = .22$, did not show any significant differences between the two groups. Therefore, it seems that video explanations may be slightly clearer than text explanations. However, given the small differences, it seems safe to conclude that there was no strong evidence that video-based interventions are more appealing or motivating than text-based interventions.

I. Direction of change by task in Study 4 (base-rate re-test), Study 5 (conjunction fallacy re-test), and Study 6 (bat-and-ball re-test)

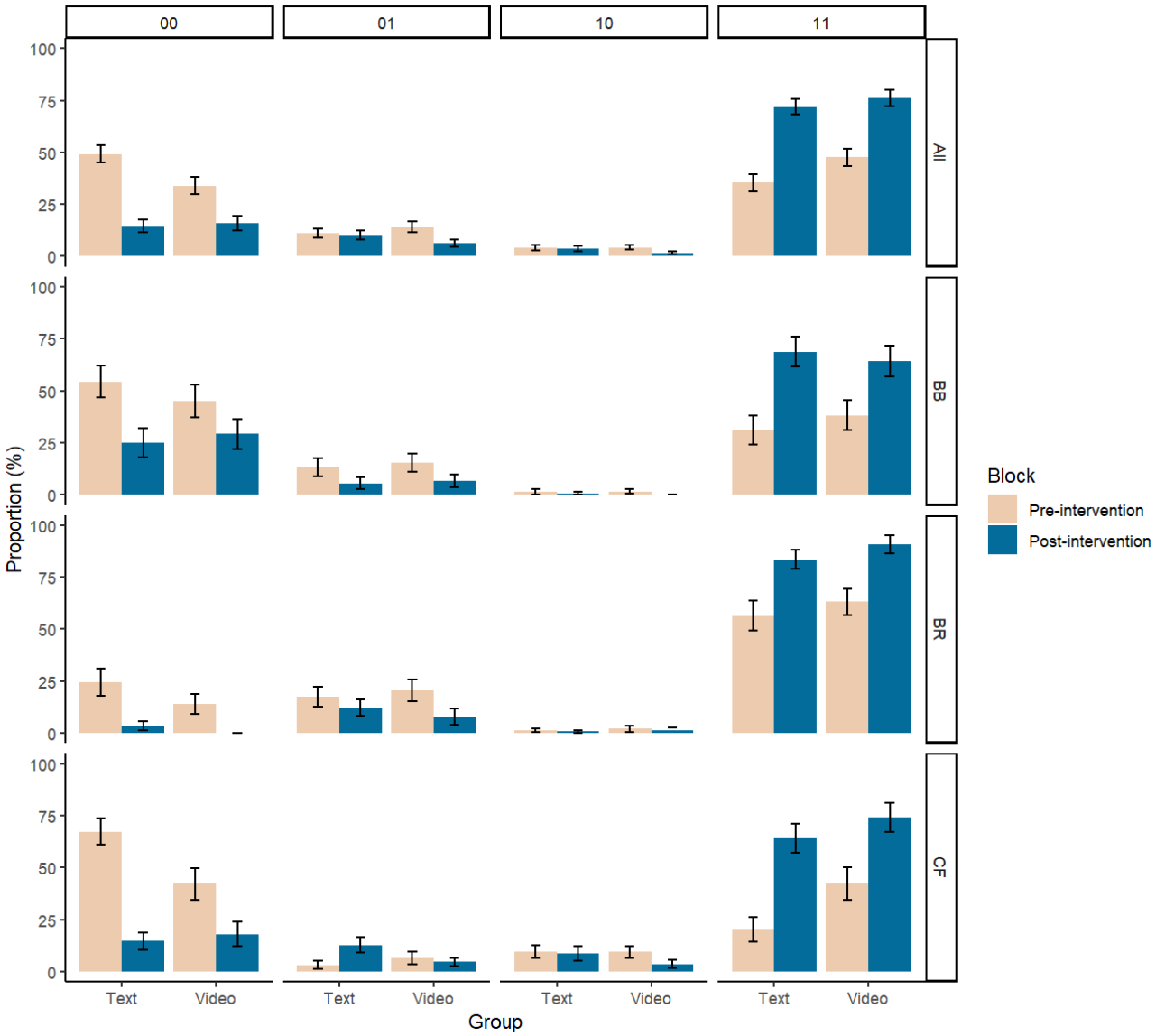

Figure S9. Proportion (%) of each direction of change (i.e., "00" pattern, "01" pattern, "10" pattern, and "11" pattern; 0 = incorrect response, 1 = correct response, first digit = initial response, second digit = final response) on conflict problems, before and after the intervention, for each task (BB, BR, CF), and combined (All) in Studies 4-6. 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.

J. Excluded trials per group and per task

Table S10. Percentage of trials excluded due to missed initial responses (before the deadline) or incorrect matrix selections in the load task, and remaining trials analysed separately for each group (video, text, and control) within each task in Studies 1-3.

		All	Video group	Text group	Control group
Study 1: Base-rate neglect	Failed deadline	1.8%	2.2%	1.7%	1.4%
	Failed matrix	13.4%	13.4%	13.9%	13.0%
	Remaining trials	85.0%	84.7%	84.6%	85.8%
Study 2: Conjunction Fallacy	Failed deadline	2.1%	2.5%	2.4%	1.4%
	Failed matrix	13.2%	11.2%	15.7%	12.5%
	Remaining trials	85.0%	86.6%	82.3%	86.3%
Study 3: Bat-and-ball	Failed deadline	1.1%	1.0%	1.4%	1.0%
	Failed matrix	14.2%	15.1%	12.6%	14.9%
	Remaining trials	84.9%	84.1%	86.2%	84.2%

Table S11. Percentage of trials excluded due to missed initial responses (before the deadline) or incorrect matrix selections in the load task, and remaining trials analysed separately for each group (video and text) within each task in Studies 4-6.

		All	Video group	Text group
Study 4: Base-rate Neglect Retest	Failed deadline	2.4%	3.1%	1.6%
	Failed matrix	10.7%	9.0%	12.3%
	Remaining trials	87.2%	88.2%	86.3%
Study 5: Conjunction Fallacy Retest	Failed deadline	3.0%	2.9%	3.1%
	Failed matrix	10.7%	7.2%	13.8%
	Remaining trials	86.6%	90.1%	83.5%
Study 6: Bat-and-ball Retest	Failed deadline	1.5%	0.9%	2.2%
	Failed matrix	11.6%	11.3%	12.0%
	Remaining trials	87.1%	87.9%	86.1%