# Examining the role of deliberation in de-bias training

Esther Boissin, Serge Caparos, Wim De Neys

# Examining the role of deliberation in de-bias training

Esther Boissin[1], Serge Caparos[2,3], Wim De Neys[1]

[1] Université Paris Cité, LaPsyDÉ, CNRS, F-75005 Paris, France

[2] Université Paris 8, DysCo lab, Saint-Denis, France

[3] Institut Universitaire de France, Paris, France

## Abstract

Does avoiding biased responding to reasoning problems and grasping the correct solution requires engaging in effortful deliberation or can such solution insight be acquired more intuitively? In this study we set out to test the impact of deliberation on the efficiency of a de-bias training in which the problem logic was explained to participants. We focused on the infamous bat-and-ball problem and varied the degree of possible deliberation during the training session by manipulating time constraints and cognitive load. The results show that the less constrained the deliberation, the more participants improve. However, even under extremely stringent conditions (high time-pressure and dual task load), participants still show a significant improvement. Critically, this "intuitive" insight effect persists over two months. This suggests that deliberation helps reasoners benefit from the training, but it is not indispensable. We discuss critical applied and theoretical implications.

***Keywords***: Reasoning; Insight; Heuristics & Biases; De-biasing; Intuition

## Introduction

Decades of research have shown that human reasoning and decision making are often biased. People tend to base their inferences on quick and intuitive impressions rather than on more costly deliberative thinking (e.g., Evans, 2008; Kanheman, 2011). In and by itself, this intuitive or so-called "heuristic" thinking can be useful because it is fast and effortless and often provides valid problem solutions. However, our intuitions sometimes cue responses that conflict with logical or probabilistic principles. One of the problems that nicely illustrates this bias is the notorious "Bat-and-ball" problem, initially presented by Frederick (2005):

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

Intuitively, most reasoners promptly conclude that the ball should cost "10 cents". However, if the ball costs 10 cents, and the bat costs $1 more, then the bat would cost $1.10. If the bat costs $1.10, then the total would be $1.20 and not $1.10 as stated. On reflection, it appears that the ball must cost 5 cents and the bat—which costs $1 more—costs $1.05.

It is striking to observe that our reasoning in the bat-and-ball problem is biased even though the solution is based on a simple algebraic equation: "X+Y=1.10, Y=1+X, Solve for X", which many adults have encountered in secondary school mathematics (Hoover & Healy, 2017). More interestingly, the “10 cents” answer is given in a majority of cases (Frederick, 2005; Toplak et al., 2014), even in samples composed of highly qualified university students (Bourgeois-Gironde & Van der Henst, 2009; Frederick, 2005), and even after repeated exposure to the problem (Raoelison & De Neys, 2019; Stagnaro et al., 2018). Although the correct "5 cents" response does not require complex mathematical operations, it seems it is not directly accessible to most people when they first come across the bat-and-ball problem (Hoover & Healy, 2017).

Sound reasoning often requires that people apply logico-mathematical rules (e.g., as the one that allows us to arrive at the "5 cents" response), which are believed to be costlier and rely on deliberate thinking, as opposed to intuitive thinking which comes quickly and effortlessly (e.g., Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996). According to the traditional view of dual-process theory, reasoning can be characterized as an interaction between these two types of thinking. This view entails that people firstly generate intuitive responses that need to be revised by the deliberative system in order to produce correct inferences (Evans & Stanovich, 2013; Kahneman, 2011; Kahneman & Frederick, 2002; Morewedge & Kahneman, 2010). Because most reasoners tend to minimize demanding computations (Kahneman, 2011), they would apply the intuitive system by default and simply stick to the response that quickly comes to mind without considering that the correct answer could be different from the intuitively generated one.

Cognitive scientists have long tried to remediate people’s biased thinking and to get them to reason correctly (e.g., Lilienfeld et al., 2009; Milkman et al., 2009). A number of recent studies have shown potential in this respect (e.g., Boissin et al., 2021; Claidière et al., 2017; Hoover & Healy, 2017; Morewedge et al., 2015; Purcell et al., 2020; Trouche et al., 2014). These “de-biasing” studies indicate that short explanations about the intuitive bias and the correct solution strategy often help reasoners produce a correct response. Once the problem has been properly explained to reasoners, they manage to solve structurally similar problems afterwards.

Moreover, available evidence suggests that these de-biasing interventions help people produce correct responses intuitively, with no need for deliberation (Boissin et al., 2021). To differentiate intuitive from deliberative responses, Boissin et al. adopted a two-response paradigm (Thompson et al., 2011), in which participants were asked to provide two consecutive answers to each problem. The initial—intuitive–response was given under time pressure and cognitive load, so as to prevent deliberation (Bago & De Neys, 2019). Immediately after giving their intuitive response, participants could take all the time they needed to give a second, final response, using deliberation. Results showed that in less than 5 minutes—after just a few explanations of the bat-and-ball problem—participants were able to provide

correct initial (intuitive) stage answers. In other words, thanks to the training, biased intuitions seemed to be replaced by correct ones (that align with logico-mathematical principles).

Nevertheless, while the study of Boissin et al. (2021) shows that a training intervention allows people to develop more correct intuitions, it does not say anything about the role of deliberation in the training process. That is, although participants were forced to intuit in the test phase after the training, they could obviously freely deliberate about the problem during the training. One may assume that deliberation is important to benefit from the training intervention, and that one first needs to actively reflect on the correct solution strategy to "see the light" and grasp it, before being able to implement it intuitively during subsequent problem solving. However, in theory, one may also consider the possibility that even during the training process, the critical problem understanding, or insight, occurs intuitively.

Interestingly, the recent literature on insight during problem solving seems to lend some credence to the latter option. It is well established that a correct solution can suddenly pop into conscientiousness typically accompanied by a so-called "Aha! Experience" (e.g., Topolinski & Reber, 2010). This sudden onset is often considered to be the result of a demanding analytic deliberation process (Chein & Weisberg, 2014) aimed at reconsidering the internal representation of the problem during which people explore all possible solutions. However, some studies question this view (e.g., Ellis et al., 2011; Stuyck et al., 2021). For example, Stuyck et al. showed that correct insight solutions still emerged under cognitive load (Stuyck et al., 2021) and do not necessarily require the engagement of deliberation.

However, while the spontaneous insight that accompanies some problem solving might have been shown to be intuitive in nature, to our knowledge no study has investigated the nature of "instructed" insight when participants are explained a solution strategy during a (debias) training. During the training, does the solution understanding or insight emerge only through the engagement of deliberation, or can it also emerge when we are forced to rely on more intuitive processing?

The current paper aimed to contrast these two possibilities. In Study 1, we studied the impact of a training intervention on bat-and-ball problem-solving performance (Boissin et al., 2021), while varying the level of available cognitive resources during the intervention, in order to impact the engagement of deliberation experimentally. While some participants received a standard training intervention in which the involvement of deliberation was not restrained ('free' training group), others received an intervention under time restriction ('fast training group'), or under both time and cognitive-load restrictions ('fast-load training group'). The benefit of the intervention on problem solving was contrasted across these three conditions and compared to the performance of a no-training control group.

In Study 2, we investigated whether the presence vs absence of deliberation during the training intervention affected the robustness of the training effect. To that effect, the trained participants of Study 1 were re-tested two months after their initial training.

# Study 1

Study 1 was run as two separate experiments. In the first experiment, there were three distinct groups: A control no-training group, a fast-training group, and a free (standard) training group, which differed only in the characteristics of the intervention. During the intervention, participants saw bat-and-ball-like problems and received an explanation after each problem on how to solve them correctly (except for the control no-training group which only saw the problems without an explanation). While the fast-training group had to read the explanations under time pressure, the free-training group could take all the time they needed to reflect on them. In a second experiment, we further constrained the potential involvement of deliberation during the fast-load training, adding cognitive load to the time constraint. Participants had to hold a pattern in memory in addition to having limited time when reading the training explanation. The data of the initial three groups (Experiment 1) and of the fourth group (Experiment 2) are presented together in the next section, in order to facilitate comprehensibility. We contrasted reasoners' performance before and after the intervention. Therefore, in a pre- and post-intervention block of trials, participants were instructed to respond either intuitively (with severe time constraints, i.e., 'fast' trials) or deliberatively (with no time constraint, i.e., 'slow' trials) to another set of bat-and-ball problems. This allowed us to assess the training impact on more intuitively ('fast' trials) vs deliberately ('slow' trials) generated responses.

## Methods

### Pre-registration and open data

The design and research questions (no specific analyses were preregistered) of experiments 1 and 2 were separately preregistered on the AsPredicted website (https://aspredicted.org) and stored on the Open Science Framework (https://osf.io/3b4jy/?view_only=a388443c8fc34310b9f908fe847f077b) where all data and material can be accessed.

### Participants

Participants were recruited online, using the Prolific Academic website (http://www.prolific.ac). Participants had to be native English speakers to take part. In total, 201 participants were tested in Study 1. Altogether, 151 individuals participated in the first experiment (107 females, M = 36.4 years, SEM = 1.2; 50 participants randomly assigned to the control no-training group, 50 to the fast-training group and 51 to the free-training group), and 50 individuals participated in the additional experiment and were assigned to the fast-load-training group (33 females, M = 33.5 years, SEM = 1.5).

In the first experiment, five participants had not completed secondary school, 58 participants had secondary school as their highest level of education, and 88 reported a university degree. In the

additional experiment, 27 participants reported secondary school as their highest level of education, and 23 reported a university degree. We compensated participants for their time at the rate of £5 per hour.

Following Boissin et al. (2021), we screened for familiarity with the original bat-and-ball problem (during the intervention, see below). In the first experiment, 26 participants reported that they already knew the problem and also provided the correct ("5 cents") response. They were excluded in order to eliminate the possibility that their prior knowledge of the correct solution would affect the results (e.g., see Bago & De Neys, 2019) and we kept 125 participants (43 in the control no-training group, 41 in the fast-training group and 41 in the free-training group). In the additional experiment, 10 participants reported having seen the bat-and-ball problem before and provided the correct ("5 cents") response. They were excluded, leaving 40 participants who performed the fast-load training in the analyses.

A sensitivity analysis indicated that our sample size would allow us to detect small to medium size training differences (effect size f = .19) between the groups with 80% power[1].

**Material**

The experiments were composed of three blocks presented in the following order: A pre-intervention, an intervention, and a post-intervention. In total, each participant had to solve 36 bat-and-ball problems (32 during the pre- and post-intervention and 4 during the intervention). Both the pre- and the post-intervention blocks always featured 8 "fast" trials, composed of 4 conflict and 4 no-conflict problems, followed by 8 "slow" trials, composed again of 4 conflict and 4 no-conflict problems (see further). During the intervention, participants were asked to respond to the classic bat-and-ball problems and three bat-and-ball-like problems that were followed by an explanation of the correct solution. Participants in the control group responded to the same three problems but were not presented with explanations. All the problems are presented in the Supplementary Material, Section A.

**Bat-and-ball problems during pre- and post-intervention.** We presented the exact same problems as in Boissin et al. (2021) taken from Raoelison and De Neys (2019). They were modified versions of the bat-and-ball problem (e.g., "In a company there are 150 men and women in total. There are 100 more men than women. How many women are there?"), which used quantities instead of prices (Bago & De Neys, 2019; Janssen et al., 2021; Raoelison & De Neys, 2019). They were presented using a free-response format, where participants typed in their response using a keyboard (e.g., see Bago & De Neys, 2019).

Half of the problems were featured in their standard "conflict" version in which the intuitively cued "heuristic" response cues an answer that conflicts with the correct answer. To ensure that participants were engaged in the task, we also presented problems which were featured in their no-conflict version,

[1] Hence, whenever we report null findings, we cannot exclude that with a more powerful design the effects might be significant.

and which were used as control problems. In these control problems, we deleted the critical relational "more than" statement. The heuristic intuition thus cued the correct response (De Neys et al., 2013; Travers et al., 2016), for instance:

*In an office, there are 150 pens and pencils in total.*

*There are 100 pens.*

*How many pencils are there in the office?*

These control problems should be easy to solve. If participants are paying minimal attention to the task and refrain from random guessing, they should show high accuracy (Bago & De Neys, 2019). Note that we added three words to the control problem questions (e.g., "How many pencils are there in the office?") in order to equate the semantic length of the conflict and no-conflict (control) versions (Raoelison & De Neys, 2019).

For each trial type (i.e., fast and slow trials) in each block (i.e., pre- and post-intervention block), two sets of eight unique problems were used (i.e., two sets during the pre-intervention for fast trials, two sets during the pre-intervention for slow trials, two sets during the post-intervention for fast trials, and two sets during the post-intervention for slow trials). The conflict problems in one set were the no-conflict problems in the other set, and vice versa. Participants were randomly assigned to one of the two sets. Consequently, none of the pre- and post-intervention problem contents was repeated within-subjects (i.e., participants saw a total of 32 different items). The presentation order of the conflict and no-conflict problems was randomized in each set.

**Fast and slow trials.** During the pre- and post-intervention, half of the trials were "fast" trials in which participants were instructed to respond as fast as possible with the first intuitive response that comes to mind. To ensure that the response was intuitive, there was a limited response time. After a fixation cross was displayed for 2 seconds, the first sentence of the problem that stated the relationship between two items (e.g., "In an office, there are 150 pens and pencils in total.") appeared for 2 seconds, followed by the entire problem (e.g., "In an office, there are 150 pens and pencils in total. There are 100 pens more than pencils. How many pencils are there?") for a maximum of 8 seconds. The time limit of 8 seconds was chosen for the fast trials, based on previous pretesting that indicated that it simply amounted to the time needed to read the preambles, move the mouse, and type an answer (see Bago & De Neys, 2019, for details). To put this in perspective, note that previous work that adopted a classic response format without time restriction indicated that participants typically need over 30 seconds to solve the bat-and-ball problem correctly (Johnson et al., 2016; Stupple et al., 2017). Hence, by all means, the 8 seconds deadline is challenging. After 6 seconds, the background of the screen turned yellow to warn participants that they only had a short amount of time left to answer. If they had not provided an answer before the

time limit, they were given a reminder that it was important to provide an answer within the time limit on subsequent trials. This reminder was displayed for a maximum of 3 seconds.

The other half of the trials were slow trials in which participants were instructed that they could take all the time they needed to reflect on the problem. Problem presentation duration was therefore unlimited for slow trials and participants could move on to the following screen by pressing the Enter key.

Note that the fast and slow trial procedure was inspired by the work of Raoelison et al. (2021a) and Markovits et al. (2019). To be clear, in contrast to the standard two-response paradigm in which reasoners give two immediately consecutive answers to each problem (one intuitive and one deliberate), they now first gave fast—intuitive—answers to a set of bat-and-ball problems, followed by slow--deliberate—answers to a second set of bat-and-ball problems (Raoelison et al., 2021a). The present design was implemented in order to prevent participants from possibly delaying their deliberation till just after the training. That is, in theory, the alternation of intuitive and deliberate trials in the standard two-response paradigm used in Boissin et al. (2021) might give participant the opportunity to deliberate about the explanations in the post-intervention block which could then artificially boost performance on the subsequent intuitive trials. Although Raoelison et al. (2021a) showed that this is unlikely, the grouping of all fast-intuitive trials before the slow-deliberation trials in the current design circumvented it completely.

**Intervention block.** The intervention block was composed of the standard bat-and-ball problem and three modified versions of it, always displayed in the same order. These bat-and-ball-like problems used prices instead of quantities, unlike in the pre- and post-intervention blocks. All participants were first shown the original bat-and-ball problem taken from Frederick (2005), for which we asked participants (1) to indicate whether they had seen this problem before, and (2) to provide an answer to the problem by typing their response and pressing 'Enter'. They had an unlimited time to respond.

Afterwards, participants saw three modified version of the bat-and-ball problem (i.e., in that order: The hat and the ribbon; "A hat and a ribbon cost $4.20. The hat costs $4.00 more than the ribbon. How much does the ribbon cost?", the banana and the apple; "A banana and an apple cost $1.40. The banana costs $1.00 more than the apple. How much does the apple cost?", and the magazine and the banana problem; "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?"). After responding to each problem, participants were explained the correct solution to apply (except in the control group), which read (e.g., for the hat-and-ribbon version):

> *The correct answer to the previous problem is 10 cents. Many people think it is 20 cents, but this answer is wrong.*
>
> *If the ribbon costs 20 cents, the hat would cost $4.20 (as it costs $4.00 more than the ribbon); both together, they would then cost $4.40.*

*However, the problem said they cost $4.20 together.*

*The correct response is that the ribbon costs 10 cents, the bat $4.10 so together they cost $4.20 ($0.10 + $4.10 = $4.20).*

The explanation was adapted from previous studies (Boissin et al., 2021; Claidière et al., 2017; Hoover & Healy, 2017; Morewedge et al., 2015; Purcell et al., 2020; Trouche et al., 2014). It was as brief and simple as possible in order to prevent fatigue or disengagement from the task. Also, the explanation provided both the correct answer and the typical incorrect answer but refrained to mention any direct heuristic mathematical shortcut such as “it is half of what you think”. To avoid promoting feelings of judgment, Boissin et al. gave no personal feedback of the type "your answer was wrong" (Trouche et al., 2014). Similarly, in order to avoid inducing mathematical anxiety, the explanation did not mention a formal algebraic equation (Hoover & Healy, 2017). Participants moved on to the following screen by clicking on the “Next” button.

We manipulated the involvement of deliberation across groups, by varying problem-solving and explanation-reading time limits, and presence or absence of additional cognitive load during the intervention. Consequently, there were four different intervention groups in this study. Participants in the free-training group had all the time they needed to answer each problem and to read each explanation, which allowed them to fully engage deliberation. In contrast, participants in the fast, fast-load, and control interventions, had a limited time to respond to each problem, and participants in the fast-training and the fast-load training groups had a limited time to read explanations. Finally, an additional cognitive load was exerted on participants in the fast-load training group.

To unify the time-restriction for explanation reading during the intervention with the problem-solving time, we set up the maximum explanation reading time at 8 seconds. A short pilot study showed that naïve subjects (n = 30, 19 females, M age = 33.3 years, SEM age = 2.1) who were simply asked to read the explanations, needed on average 19 seconds (SEM = 2) to read each explanation. This indicates that the 8 seconds deadline, which is less than the fastest quintile of the pilot reading times, was highly challenging.

The cognitive load task used in the fast-load intervention, was based on the dot memorization task (Miyake et al., 2001), given that it had been successfully used to burden executive resources in previous reasoning studies (e.g., De Neys, 2006; Franssens & De Neys, 2009). Participants had to memorize a complex visual pattern (i.e., 4 crosses in a 3x3 grid) presented during 2 seconds before each explanation. After having read the explanations, participants were shown four different patterns and had to identify the one that they had to memorize (see Bago & De Neys, 2019, for more details). Participants had a maximum of 6 seconds to make their selection.

**Equation solving.** For exploratory purposes, after the last conflict problem in the post-intervention, we asked participant to choose the equation which best described how they arrived at their response. They were given two possible choices. For example:

*At a convention there are 560 neuroscientists and botanists.*

*There are 500 more neuroscientists than botanists. How many botanists are there?*

*Imagine you would have to explain how you solved the previous problem in mathematical terms.*

*Which one of following equations describes best how you arrived at your answer?*

*(x stands for botanists)*

- *500 + x = 560*
- *500 + 2x = 560*

Mathematically speaking, in this example the equation leading to the correct answer is "500 + 2x = 560". However, biased participants tend to use the simplified equation: "500 + x = 560" (Kahneman, 2011; Kahneman & Frederick, 2002). This equation solving procedure was meant to verify whether participants who had benefited from the training intervention had identified the correct logico-mathematical structure of the bat-and-ball problems. By and large, reasoners who were explained how to correctly solve bat-and-ball-like problems more frequently chose the correct equation than those who received no training. This suggests that once participants were presented the explanation, they tend to recognize the logico-mathematical structure of the problems. However, these equation solving data were collected for exploratory purposes and should be interpreted with caution (data and statistical analyses are presented in the Supplementary Material Section B).

**Bat-and-two-balls problems.** At the end of both pre- and post-intervention blocks, participants from the fast-load-training group were asked to respond to two "bat-and-two-balls" problems taken from Boissin et al. (2021):

*A bat and two balls cost $2.60 in total.*

*The bat costs $2 more than two balls.*

*How much does one ball cost?*

These problems aim to test for a possible heuristic confound. That is, it is possible that our explanations do not help to clarify the underlying logic but simply let participants develop a new heuristic (e.g., "it's half of what you think it is!"). Although our no-conflict control problems should allow us to identify such a blind "halving heuristic", they nevertheless have a different underlying structure (e.g., they do not contain "*more than X*") that might be used as an advanced selective halving cue. Thus, with bat-and-two-balls problems, we wanted to have some additional control. Specifically, this problem shares the same basic underlying logic as the original bat-and-ball problem. Contrary to the no-conflict control problems, it contains the "more than" statement which leads to the emergence of a heuristic response

("60 cents") that conflicts with the correct response ("15 cents"). The difference with the original bat-and-ball is that it specifies the relation between three objects (e.g., a bat and TWO balls). Mathematically speaking, the following equation needs to be applied in order to solve bat-and-two-balls problems: "Y + 2X = $2.60. Y = $2 + 2X; or 4X = $.60" vs traditional bat-and-ball structure: "Y + X = $2.60. Y = $2 + X; or 2X = $.60". Hence, reaching the correct response ("15 cents") requires an additional division. Critically, however, the basic equation substitution logic is completely similar. Accordingly, the bat-and-two-balls problem elicits three types of responses: Heuristic (x = $2.60 - $2), halving (x = ($2.60 - $2) /2), and correct (x = ($2.60 - $2) / 4). If reasoners simply use the blind halving strategy ("30 cents") after the intervention, they will err, but if they understand the bat-and-ball logical structure, then in theory they should also manage to solve the bat-and-two-balls problem. Boissin et al. (2021) showed that participants who received an intervention similar to that of the free intervention group, i.e., without any time or cognitive constraints, produced more correct responses to bat-and-ball and bat-and-two-balls problems after the intervention. Thus, the idea of presenting these problems was to explore whether participants who improved under restrictive-training conditions also showed this generalization towards the bat-and-two-ball problems. Overall, the results showed that the fast-load intervention also improved participants' accuracy at the bat-and-two-balls problems, suggesting that they understood the logico-mathematical principles underlying the bat-and-ball problems. Clearly, this problem was added for exploratory purposes and the results should consequently be interpreted with caution again (data and statistical analyses are presented in the Supplementary Material Section C).

**Procedure.** The experiment was run online using the Qualtrics platform. Participants were instructed that the experiment would take twenty minutes and that it demanded their full attention. A general description of the task was presented in which participants were instructed that they would need to solve reasoning problems either quickly or slowly. They were told that we were interested either in their very first, initial answer that comes to mind, with no reflection involved, or in a slow, reflective answer, when one has all the time to deliberate. In order to familiarize the participants with the response deadlines, they solved two fast practice trials, with time constraint, followed by two slow practice trials, with no time constraint. At the end of the practice, we clarified that the trial types (fast or slow) would be grouped in sets of eight trials, and that participants would be informed about trial type before each set. To visually remind participants of which type they were solving, instructions for the fast trials were presented in green font, and in blue font for the slow trials. For the fast trials, the screen turned yellow after 6 seconds (i.e., 2 seconds before the deadline) to visually remind participants that the deadline was approaching. After the pre-intervention, which consisted of 8 fast and 8 slow trials, participants moved to the intervention block. Depending on the group they were assigned to, participants were informed that they would (control, fast, and fast-load intervention groups) or would not (free group) have a time constraint

to solve the intervention problems. Participants in the relevant groups (fast, fast-load, and free training group) were further informed that they would be given an explanation after each of their responses, that there would be a time constraint of 8 seconds to read the explanations (fast and fast-load groups) and that there would be an additional load memorization task to perform (fast-load group). Participants in the fast-load group also were given two memorization task practice trials in which they practiced the memorization task only. They were also familiarised with the entire intervention procedure composed of the memorization task and the reading of an unrelated practice text.

In the fast group, we opted for a fixed (limited) explanation reading time: The explanations were always presented for the maximum allotted time of 8 seconds. In the fast-load group we allowed participants to freely advance in case they finished reading before the 8 seconds deadline. Note that the average explanation reading time in the fast-load-training group was very closed to 8 seconds (see below) suggesting that 8 seconds as a fixed deadline was appropriate for the fast-training-group.

At the end of the experiment, participants from the control no-training group were presented with the explanations about how to solve the bat-and-ball problems, and all participants were asked to complete a page with demographic questions.

**Trial exclusion.** We discarded 1.8% trials in which participants failed to provide their fast answer before the deadline during the pre- and post-intervention blocks and we analysed the remaining 98.2% of all experiment trials. On average, each participant contributed 31.7 (SEM = 0.1) trials (out of 32). During the intervention, all participants succeed at providing a response to each problem before the deadline. On average, participants in the fast-load training group memorized the correct pattern in 79.2% (SEM = 3.7, range from 33.3% to 100%) of the explanation trials. This high accuracy indicates that the load task was overall properly performed. Note that for calculation of the explanation reading time (see below) we discarded the trials in which the load recall was incorrect (i.e., 20.8% of the intervention trials). On average, each participant contributed 2.6 (SEM = 0.1) trials during the intervention out of 3.

**Statistical analyses.** The data were processed and analysed using the R software (R Core Team, 2017) and the following packages (in alphabetical order): *ez* (Lawrence & Lawrence, 2016), ggplot*2* (Wickham, 2009), and *tidyr* (Wickham & Wickham, 2017).

## Results

**Bat-and-ball response accuracy.** For each participant, we calculated the average proportion of correct fast and slow responses for the conflict problems, in the pre and post intervention blocks. First, we focus on accuracies for the fast responses (i.e., for the first eight problems of the pre- and post-intervention blocks). Figure 1 shows that reasoners, in all groups, typically failed to produce correct responses to

conflict problems before the intervention but improved after intervention. A mixed-design ANOVA on the fast response accuracy with Block (Pre- vs Post-intervention) as a within-subject factor and Group (Control vs Fast vs Fast-load and Free training-groups) as a between-subject factor, showed that there was an overall improvement after the intervention (main effect of Block; $F(1,161) = 147.59$, $p < .001$, $\eta^2g = .28$) that varied across groups, the Block x Group interaction reached significance; $F(3,161) = 8.03$, $p < .001$, $\eta^2g = .06$. Follow-up 2 x 2 mixed-design ANOVAs with Block as a within-subject factor and Group as between-subject factor, showed a larger improvement for each of the different training groups than in the control no-training group, that is the Block x Group interaction for each comparison showed significance (Free vs Control: $F(1, 82) = 22.8$, $p < .001$, $\eta^2g = .13$; Fast vs Control: $F(1, 82) = 12.9$, $p < .001$, $\eta^2g = .06$; and Fast-load vs Control: $F(1, 81) = 8.03$, $p < .001$, $\eta^2g = .05$). In addition, reasoners who were allowed to freely deliberate during the intervention (i.e., the free-training group) showed a higher performance increase than reasoners in the fast-training group, The Block x Group interaction reached significance; $F(1, 80) = 1.75$, $p = .02$, $\eta^2g = .01$, and than reasoners in the fast-load one, the Block x Group interaction reached significance; $F(1, 79) = 3.52$, $p < .001$, $\eta^2g = .02$. Finally, despite the visual trend observed in Figure 1, there was no significant difference between pre-to-post performance improvement in the fast-training group and in the fast-load-training group; the Block x Group comparison of Fast and Fast-load groups accuracies showed no significant interaction, $F(1, 79) = 0.32$, $p = .57$, $\eta^2g = .002$.

Second, we focus on accuracies for the slow responses (i.e., for the last eight problems of the pre- and post-intervention blocks). Figure 1 shows that reasoners, in all groups, typically failed to produce correct slow responses to conflict problems before the intervention but improved after the intervention. A mixed-design ANOVA on the slow response accuracy with Block (Pre- vs Post-intervention) as a within-subject and Group (Control vs Fast vs Fast-load vs Free-training-groups) as a between-subject factor showed that there was an overall improvement after the intervention (main effect of Block; $F(1,161) = 73.80$, $p < .001$, $\eta^2g = .10$) that varied across groups, the Block x Group interaction reached significance; $F(3,161) = 4.16$, $p = .01$, $\eta^2g = .19$. Follow-up tests using 2 x 2 mixed-design ANOVAs with Block as a within-subject factor and Group as between-subject factor, showed that overall, the improvement was always larger in each of the different training groups than in the control no-training group, the Block x Group interaction was significant for each comparison (Free vs Control: $F(1, 82) = 13.02$, $p < .001$, $\eta^2g = .04$; Fast vs Control: $F(1, 82) = 6.43$, $p = .001$, $\eta^2g = .02$; and Fast-load vs Control: $F(1, 81) = 3.49$, $p = .06$, $\eta^2g = .001$). The follow-up 2x 2 ANOVA tests further indicated that there was no strong evidence for a differential improvement among the different training groups on the slow trials. The Block x Group interactions failed to reach significance (Free-training vs Fast-training groups: $F(1, 80) = 1.17$, $p = .28$, $\eta^2g = .004$, Free-training vs Fast-load-training group: $F(1, 79) = 2.64$, $p = .11$, $\eta^2g = .01$, and, Fast vs Fast-load training group, $F(1, 79) = 0.33$, $p = .57$, $\eta^2g < .001$).

Finally, we analysed the performance for the no-conflict control problems. As expected, it was consistently at ceiling, with grand means of 90.2% (SEM = 1.3) for fast trial accuracies, and 95.5% (SEM = 0.9) for slow trial accuracies (See Supplementary Material Section D). In line with previous studies (Bago & De Neys, 2020; Pennycook et al., 2015; Raoelison & De Neys, 2019), participants' high accuracy rates on the no-conflict problems indicated that they were paying attention to the task and refrained from random guessing.

To sum-up, explaining how to solve bat-and-ball problems led reasoners to significantly enhance their performance, compared to reasoners who received no explanations. This improvement was observed both when deliberation was minimized during reasoning (fast trials) and when it was allowed (slow trials). This replicates Boissin et al.'s finding that after a debiasing training reasoners manage to intuit the correct problem solution. Critically, the more deliberation could be engaged during the training, the more participants tended to improve their performance after the training, both in fast and slow trials. However, at the same time, even under our most stringent test conditions in which participants received the training under time pressure and load, performance still significantly improved. This suggest that although deliberation boosts training, learning can–to some extent–still be observed even when deliberate reflection is minimized.

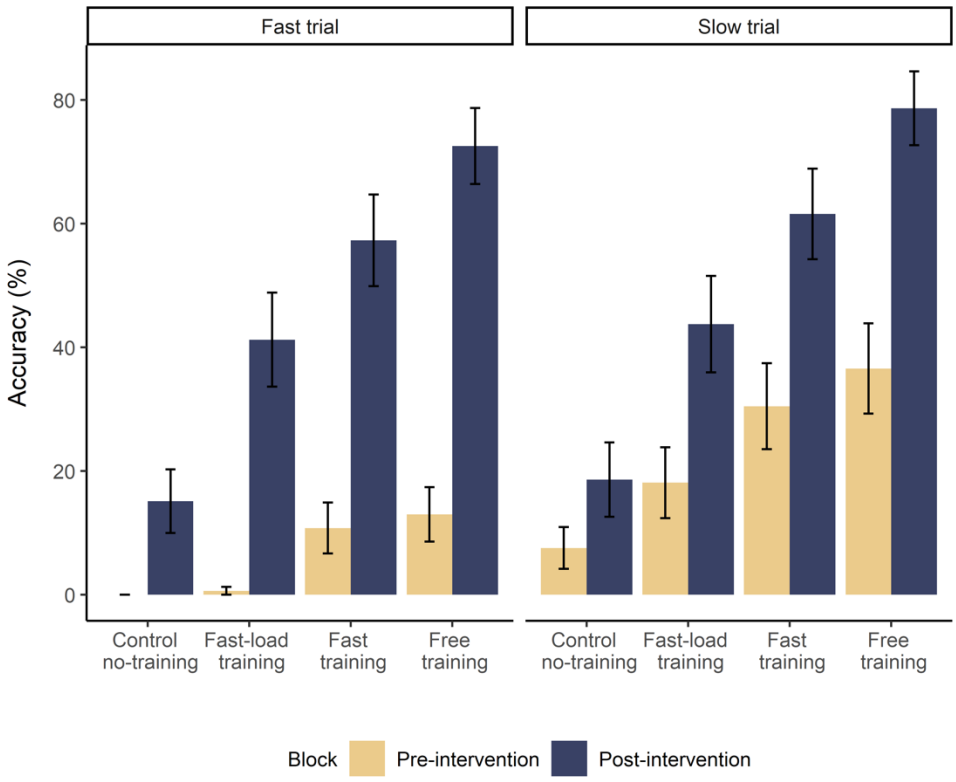


**Figure 1.** Average fast and slow trial accuracy on conflict problems in Study 1, for each group, before and after the intervention. Error bars represent standard errors of the mean (SEM).

**Individual training effect classification.** To explore further how participants benefited from the training (or not), we classified reasoners according to an individual level accuracy analysis for each participant, on each conflict trial, from start to end of the experiment (see Figure 2). We created three different categories distinguishing participants who improved after the training intervention from those who did not improve and from those who already showed accurate reasoning performance before the intervention. The categories were created according to the number of correct and incorrect responses for each type of trial (i.e., fast, and slow trials) in each block (i.e., pre-, and post-intervention). Classification was primarily based on majority of given responses, except for cases where participants provided an equal number of correct and incorrect responses. In such cases, the response to the last trial was used for classification. Thus, participants who gave more incorrect than correct responses in a specific trial type and block were classified as '0'. Participants who gave more correct than incorrect responses were classified as '1'. Hence, each participant was represented by a combination of four classification indices, one for each trial type in each block. For example, a participant with the classification '00-11' gave a majority of incorrect responses in both fast and slow trials before the intervention, but a majority of correct responses in both types of trials after the intervention. Based on these combinations of classification indices, we established the three distinct categories for participants. Reasoners classified as '11-11' and '01-11' were categorized as “correct” reasoners. Reasoners classified as '00-11' and '00-01' were categorized as “improved” reasoners. All other reasoners who did not fall into these categories were classified as "biased".

First, 16% of all participants in the free-training group, 37% in the fast-training group, 58% in the fast-load-training group and 81% in the control group were classified as “biased”. Second, participants who did not require any training intervention to respond correctly to the bat-and-ball problems and were labelled as “correct” reasoners represent 39% of the participants in the free-training group, 32% in the fast-training group, 20% in the fast-load-training group and 9% in the control no-training group. Third, improved reasoners (i.e., participants who gave a majority of incorrect responses before the intervention, at the fast and slow trials, but a majority of correct ones after the intervention for the fast and slow trials or only for the slow trials) represent 44% of participants in the free-training group, 32% in the fast-training group, and 23% in the fast-load-training group. 9% of the participants in the control group showed a spontaneous improvement in the absence of any explanation. The average proportion of correct post-intervention fast trials among the improved reasoners was consistently high (Free-training group: 82%, Fast-training group: 79% and Fast-load-training group: 81%). This indicates that for those reasoners who benefitted from the intervention, limiting deliberation during the intervention did not prevent them from automating the correct solution strategy.

In sum, in line with the overall results, the individual level analysis indicates that, limiting the possibility to deliberate decreased the impact of the training: 72% of biased reasoners prior the

intervention improved in the free training group, while the proportions in the fast and fast-load-training groups were lower, respectively 46% and 28%. Hence, the more we prevented participants from engaging in deliberation, the fewer of them benefited from the intervention. This establishes that to benefit from the training, deliberation during the intervention is clearly useful. However, at the same time, even in the most challenging conditions, up to almost a third of reasoners were nevertheless able to benefit from it. This improvement already manifested itself on the fast trials. Hence, although fewer participants benefitted from the restricted training, those who benefitted learned equally well and showed similar levels of proficiency. This suggests that, once participants have understood how to achieve the correct solution through the explanations, they are, in majority, able to apply them quickly and effortlessly, regardless of the restrictions applied during the intervention.

**Manipulation check.** Participants in the free training group did not face time or load constraints while reading the intervention explanations and solving associated training problems. We assumed that this group would deliberate about the intervention explanations and problem solutions. To verify this, we contrasted the average explanation reading and training problem solution time[2] during the intervention in the free, fast, and fast-load training groups. All latencies were log transformed prior to statistical analysis. Figure 3 reports average back-transformed values for ease of interpretation. It is readily clear that the free training group took considerably more time to read and complete the intervention problems and explanations. A one-way ANOVA showed that the average explanation reading, $F(2, 119) = 2.75$ , $p = .07$, $\eta^2g = .04$, and problem solving, $F(2, 119) = 15.31$, $p < .001$, $\eta^2g = .20$, times indeed differed (marginally) significantly between the groups. This indicates that "free" participants were indeed more likely to engage in a deliberative processing during the intervention. Following our pretesting, this also establishes that participants in the fast and fast-load groups were responding under time pressure.

Nevertheless, critics could argue that even our most challenging test condition, the fast-load group (in which participants faced both time-pressure and load during the intevention) might not have prevented all possible deliberation. That is, at least some reasoners might still have taken the time to engage in deliberation and it might be those who are driving a potential remaining training effect. Given that deliberation is slower than intuiting, such resdiual deliberation should be reflected in longer intervention times. We therefore computed correlations between intervention times (both for the problem-solving and the reading explanation times) and post-intervention accuracy for fast and slow responses, for the participants in the fast-load-training group[3].

[2] Participants solved three training problems and read an explanation for each one. We focused on the problem solution times for the two problems after the first explanation (i.e., the third and fourth problem of the whole intervention).

[3] Given that all reading times were fixed at 8 seconds in the fast group (i.e., participants could not advance earlier), this analysis was not informative in this group.

Results showed that the correlations were small and tended to be negative (intervention problem-solving latencies and post-intervention accuracy in fast, $r(38) = -0.15$, $p = .36$, and slow trials, $r(38) = -0.06$, $p = .73$, reading explanations times and post-intervention accuracy in fast, $r(38) = -0.28$, $p = .08$, and slow trials, $r(38) = -0.23$, $p = .16$). Hence, if anything, subjects who tended to take longer on the intervention were less likely to benefit from it. This argues against the claim that the training effect in the fast-load group results from residual "slow" deliberation per se.

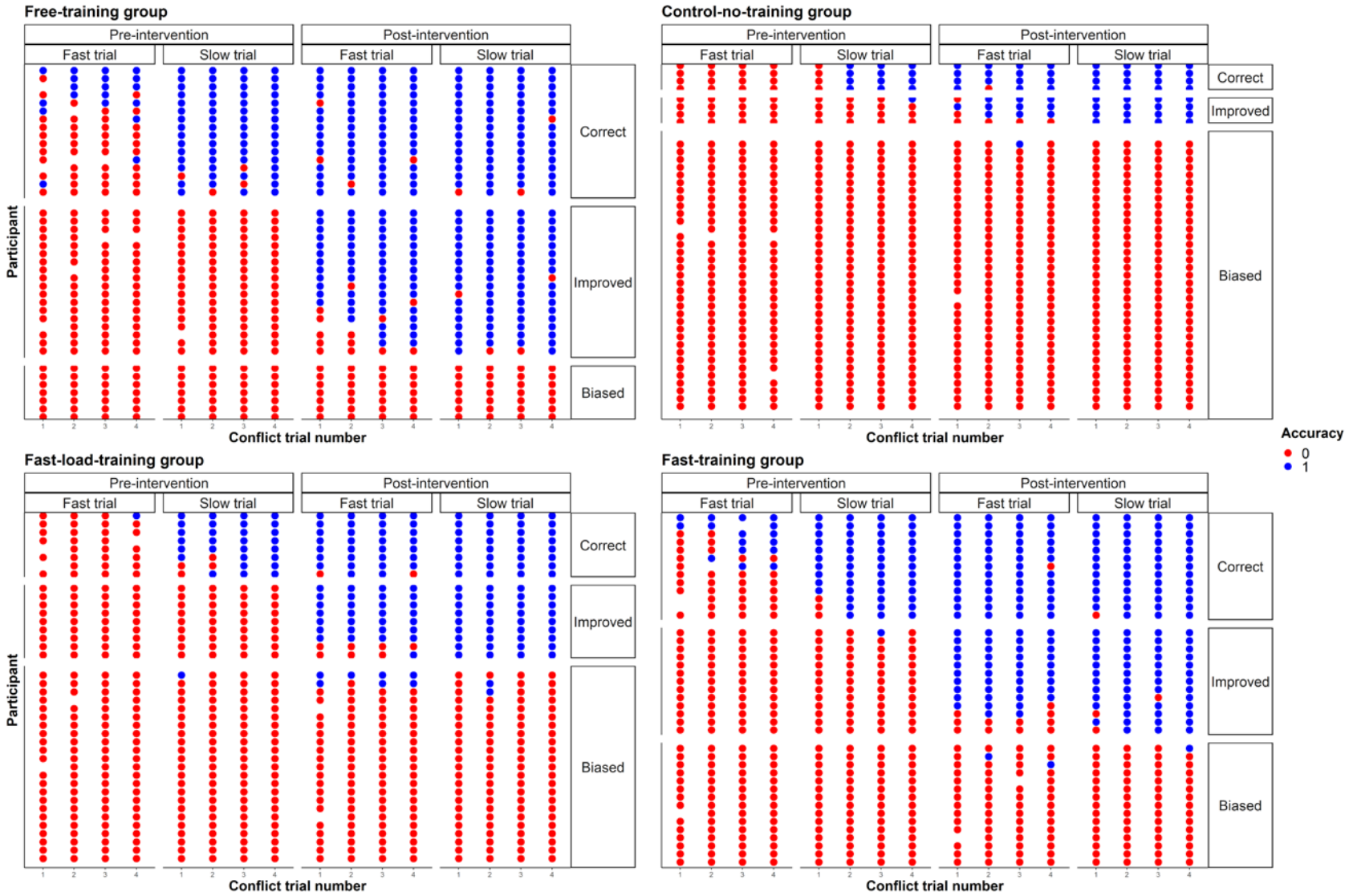


**Figure 2**. Individual training effect classification (each row represents one participant) of Study 1 both for the fast and slow trials. Due to the discarding of missed deadline trials in the fast trials (see Trial Exclusion), not all participants contributed 16 analysable trials.

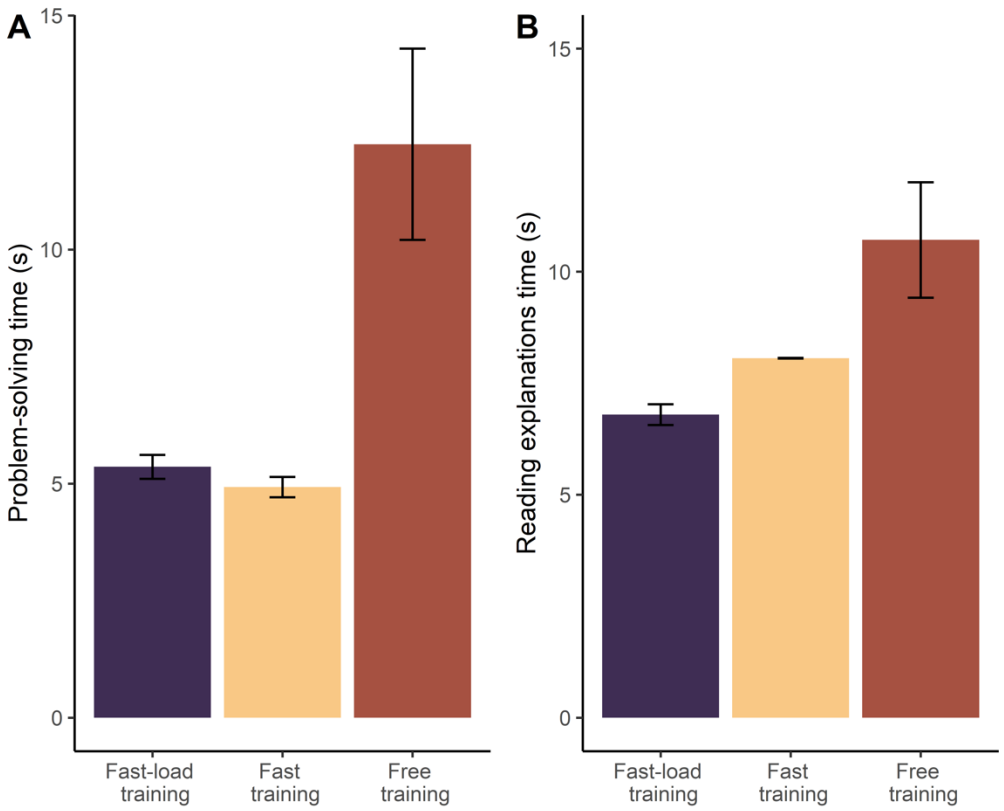


**Figure 3.** Intervention problem-solving mean time in panel A and explanation-reading mean time in panel B for each training group during the intervention. Error bars represent standard error of the mean (SEM).

## Study 2

Study 1 showed that trained participants managed to intuit the correct solution after the training intervention. Nevertheless, when deliberation was minimized during the intervention, the training effect decreased, and fewer participants benefitted from it. However, even under extremely challenging training conditions, participants were able to improve their performance as early as the intuitive fast trials. In Study 2 we tested the robustness of this training effect. Boissin et al. (2021) already established that their (free) training effect persisted up to two months after the training. It is possible that the training effect under restricted deliberation is more superficial. That is, although the more intuitive restricted training might still allow participants to grasp the solution and allow them to apply this instantly, the comprehension might be less elaborate or profound and therefore would not persist long after the training. In Study 2, we asked the trained participants from Study 1 to take part in a re-test, two months after the intervention, in order to test for the potential long-term sustainability of the training effect.

## Methods

### Pre-registration and open data

The design and research questions (no specific analyses were preregistered) were preregistered on the AsPredicted website (https://aspredicted.org) and stored on the Open Science Framework (https://osf.io/3b4jy/?view_only=a388443c8fc34310b9f908fe847f077b) where all data and material can also be accessed.

### Participants

We managed to recruit 90 participants to take part in Study 2 (out of the 122 trained participants of Study 1; 61 females, M = 37.7 years, SEM = 1.3) of whom 30 belonged to the free training group, 31 to the fast-training group, and 29 to the fast-load-training group. We compensated participants for their time at the rate of £7 per hour.

### Materials & Procedure

The material and the procedure were the same as in the pre-intervention block of Study 1. All the problems featured modified contents (see Supplementary Material Section A).

**Trial exclusion.** We discarded trials in which participants failed to provide their fast answer before the deadline (2.9% of all trials) and we analysed the remaining 97.1% trials. On average, each participant contributed 15.8 (SEM = 0.5) trials out of 16 in Study 2.

## Results

To test whether the training effect sustained over time, we compared pre- and post-intervention performance in Study 1 to performance in Study 2 (i.e., two months later), for each intervention group.

First, we focus on accuracies for fast responses. Figure 4 shows that, in all groups, reasoning accuracy decreased two months after the intervention (i.e., Study 2 retest vs post intervention Study 1). A mixed-design ANOVA on the fast response accuracy with Block (Study 2 retest vs post-intervention Study 1) as a within-subject factor and Group (Fast vs Fast-load vs Free training-groups) as a between-subject factor, showed an overall accuracy decrease two months after the training (main effect of Block, $F(1, 87) = 52.22$, $p < .001$, $\eta^2g = .14$) that did not vary across training groups, the Block x Group interaction failed to reach significance: $F(2, 87) = 1.07$, $p = .35$, $\eta^2g = .001$. Critically, performance remained higher two months after the intervention than before in all groups. A mixed-design ANOVA on the fast response accuracy with Block (Study 2 retest vs pre-intervention Study 1) as a within-subject factor and Group (Fast vs Fast-load vs Free training-groups) as a between-subject

factor, showed higher performance two months after the training than before (main effect of Block, $F(1, 87) = 18.64$, $p < .001$, $\eta^2g = 0.07$). This higher performance two months after the training did not vary across training groups, the Block x Group interaction failed to reach significance: $F(2, 87) = 1.56$, $p = .22$, $\eta^2g = 0.01$. This indicates that the training effect on fast trials after a delay of two months sustained equally well for all intervention groups.

The same pattern was found for the slow-trial responses. Performance decreased two months after the intervention (main effect of Block, $F(1, 87) = 21.04$, $p < .001$, $\eta^2g = 0.04$), to the same extent in all intervention groups. The Block x Group interaction failed to reach significance, $F(1, 87) = 1.22$, $p = .30$, $\eta^2g = 0.00$. Similarly to fast trials, slow trials performance remained higher two months after the intervention than before (main effect of Block, $F(1, 87) = 22.38$, $p < .001$, $\eta^2g = 0.05$) with no significant difference between groups, that is the Block x Group interaction failed to reach significance: $F(2, 87) = 0.34$, $p = .71$, $\eta^2g = 0.00$.

In sum, for both fast and slow trials, the training effect sustained over time and there was no difference in sustainability across types of interventions (free or deliberation-restricted). Once reasoners were trained and managed to correctly solve bat-and-ball-like problems, they remained able to intuitively apply the correct solution strategy two months after the training, even when the training occurred with severe deliberation restrictions. Hence, there is no strong evidence suggesting that restricted deliberation during the intervention led to a less profound or robust training. Note that 74% (90/122) of the trained participants from Study 1 participated in Study 2. To check for a possible attrition confound (e.g., subjects who did better in Study 1 were more likely to sign-up for Study 2), we compared the Study 1 pre-intervention conflict problem accuracy of the participants who took part in the re-test (Fast trial accuracy: M = 8.8%, SEM = 2.5; Slow trial accuracy: M = 26.7%, SEM = 4.4) to that of the participants who did not take part (Fast trial accuracy: M = 6.5%, SEM = 3.6; Slow trial accuracy: M = 33.6%, SEM = 8.14). Given that there were no clear systematic differences across these two groups of participants (Fast trials: $t(120) = 0.48$, $p = .63$, $d = .10$; Slow trials: $t(120) = 0.78$, $p = .44$, $d = .16$), it is unlikely that the Study 2 results are artificially boosted because of an attrition confound.

For completeness, we also analysed no-conflict problem accuracies. For both fast and slow trials, performance remained near ceiling for all intervention groups (see Supplementary Material Section D).

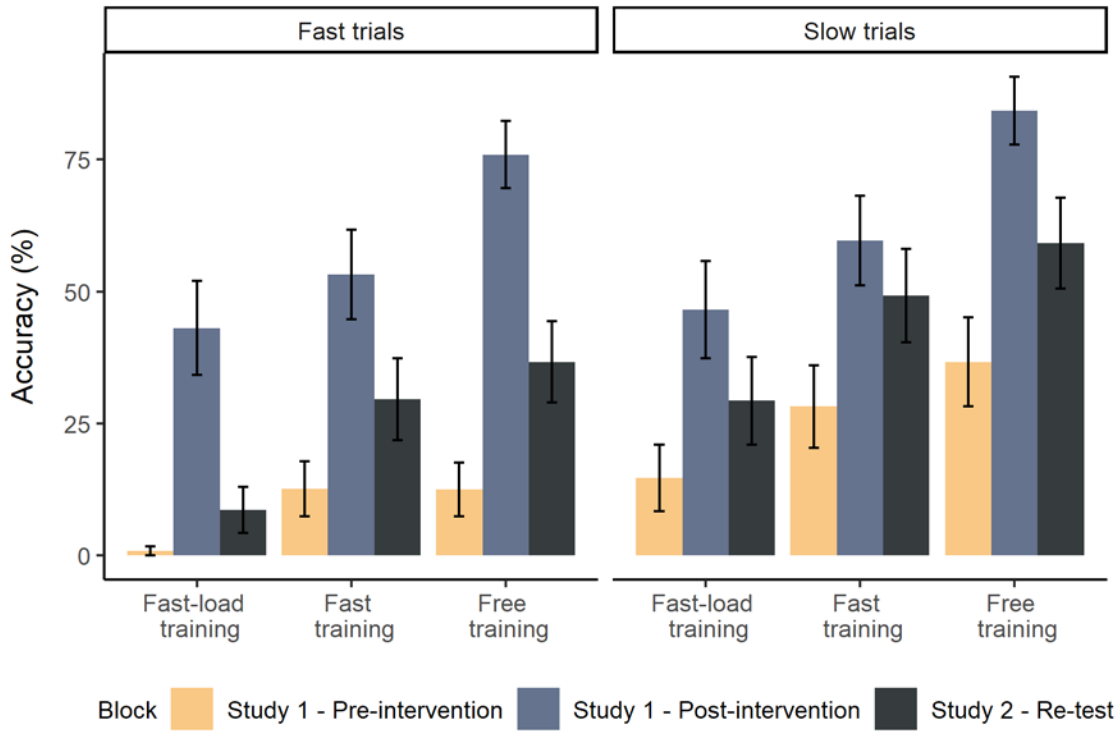


**Figure 4.** Average fast and slow trial accuracy on conflict problems in Study 1 (pre- and post-intervention) and re-test Study 2. Errors represent standard errors of the mean (SEM).

## General discussion

Does avoiding biased responding to reasoning problems and grasping the logical problem solution requires engaging in slow and effortful deliberation or can such solution insight be acquired more intuitively? The present study addressed this question and tested the impact of deliberation on the efficiency of a de-bias training in which the problem logic was explained. We focused on the infamous bat-and-ball problem and varied the degree of possible deliberation during the training session by manipulating time constraints and cognitive load. Results point to three key findings. First, in line with previous training studies, we observe across all our training groups that training improved performance. This improvement was not only observed on "slow" trials in which participants were able to deliberate but also on "fast" trials in which time-pressure minimized deliberation. This suggests that, after training, participants manage to automatize the correct solution strategy to some extent and apply it effortlessly. Hence, their "fast" intuitive hunches were correct and no longer needed to be corrected by slower deliberate reasoning (Boissin et al., 2021).

Second, although we observed some improvement across all our training groups, results clearly indicate that deliberation during the intervention boosts the training effect. The less we restrained deliberation when the problems were explained to participants, the higher the average accuracy boost after training and the more individuals benefited from it. This directly points to the beneficial role of

deliberation in generating solution insight and helping people to grasp the underlying problem logic. The more time people take to read and reflect on the intervention explanations and practice problems, the more likely that they will manage to solve the problems correctly afterward.

Third, although deliberation boosted the training effect, even under our most challenging test conditions (time-pressure and dual task load in our "fast-load" group) that restricted possible deliberation during the intervention, participants still showed a significant improvement. Critically, although fewer participants benefitted when deliberation was restricted, those who did benefit learned equally well and showed similar levels of proficiency as participants in the unrestricted training group. Hence, once participants understand how to achieve the correct solution with the help of the explanation intervention, they are, in majority, able to apply the solution quickly and effortlessly, regardless of how much they actually reflected on the explanations. This lends credence to the idea that problem solution insight can also occur more "intuitively" in the absence of deep deliberation. In other words, what our findings show is that although deliberation helps reasoners benefit from the training, it is not necessarily indispensable to get people to understand the problem logic.

The present findings may have important applied and theoretical implications. At the theoretical side they provide further insight into the nature of sound reasoning and reasoning bias. Note that even the unrestricted "free" training is fairly minimalistic. It takes no more than 5 minutes and simply illustrates the correct principle with a few examples. The fact that this nevertheless suffices to remediate biased reasoning for bat-and-ball problem already suggests that the bias does typically not result from a lack of knowledge or so-called "mind" or "storage" gap per se (e.g., Boissin et al., 2021; De Neys & Bonnefon, 2013; Hoover & Healy, 2017, 2019; Stanovich, 2011). Obviously, it is unlikely that a five-minute deliberation suffices to learn the underlying logico-mathematical principles ex nihilo. The present study strengthens this conclusion by showing that the training not even necessarily requires extensive deliberation. Bluntly put, if people can grasp the solution strategy when they barely have the time to read the explanation and are focusing their attention on another task, this indicates that solving the problem cannot be "hard." That is, the restricted training presumably worked because the critical knowledge to solve the problem was already implicitly there. Reasoners simply needed to be reminded—even if only superficially—of its relevance. Once reasoners are given this slight push in the right solution direction, grasping it can become a "no-brainer."

As Boissin et al. (2021), we believe that the current findings may also fit with recent evolutions in dual process theorizing (De Neys, 2017; De Neys & Pennycook, 2019). Traditionally, reasoning in line with logical principles on classic bias tasks is assumed to require a deliberate correction process in which "System 2" reflection overrides an erroneous initial "System 1" intuition (Evans & Stanovich, 2013; Kahneman, 2011). Hence, reasoning correctly is believed to require demanding deliberation. Recent "dual process models 2.0" have questioned this assumption (e.g., De Neys, 2017; De Neys &

Pennycook, 2019; but see also Ghasemi et al., 2022; Thompson et al., 2018). The idea is that bias tasks will evoke diverse types of intuitions. One of these might be based on erroneous "heuristic" associations but another intuition can be based on elementary knowledge of the logical principle that is evoked in the task. It is hypothesised that throughout the school curriculum, people automatize the application of basic logico-mathematical operations to some degree (e.g., De Neys, 2012; Raoelison et al., 2021b; Stanovich, 2018). However, both intuitions would have a different activation strength. Typically, for most reasoners the heuristic solution will dominate the logical one and lead to biased responding. In light of this framework, one could argue that the training helps to boost the activation of the logical intuition. By stressing the relevance of the logical problem solution, its activation level will increase and can dominate the competing heuristic intuition. Consequently, when people are faced with the same task afterward, the logical solution can be favoured even without any further deliberation.

Note that dual process models have also started to redefine the role of deliberation in reasoning (De Neys, 2017). Rather than being necessary for the correction of erroneous intuitions per se, it is assumed that deliberation might be primarily used to find explicit justifications and arguments to support intuitively generated responses (Bago & De Neys, 2019; see also Evans, 2019). For example, Bago and De Neys (2019) observed that after deliberation (in a two-response study) people had little trouble properly justifying their logical responses. Such correct justifications were much less likely for "intuitive" logical responses in the initial response stage. Hence, reasoners would need to engage in deliberation to help them come up with a proper, explicit justification for their intuitive logical insight. Such a justification or argumentation process will obviously be critical for communication and cultural knowledge transmission (e.g., Mercier & Sperber, 2011, 2017). Once a group member has found a solution to a problem, a good justification will help them to convince others to adopt it (Claidière et al., 2017; Mercier & Claidière, 2022; Trouche et al., 2014). The current finding that deliberation helps reasoners grasp the problem solution during training lends credence to this view. That is, the fact that people are more likely to adopt a solution justification when they can deliberate about it supports the idea that deliberation might be especially useful in an argument evaluation process. While speculative, this broader theoretical framework presents a potential mechanism to make sense of the current findings.

We believe that our study can also be relevant beyond the dual process and reasoning bias field per se. We noted that the recent literature on the "Aha! Experience" (Topolinski and Reber, 2010) and spontaneous insight during problem solving started questioning the assumption that such insight necessarily requires demanding deliberation (e.g., Ellis et al., 2011; Stuyck et al., 2021). Spontaneous correct insight solutions still emerge under cognitive load, for example (Stuyck et al., 2021). The present findings indicate that a similar conclusion might hold for *instructed* problem insight or

understanding. People are more likely to grasp a problem solution that is explained to them when they can deliberate but such deliberation is not indispensable. Although both research traditions clearly have their differences, the similarity between the conclusion gives some credence to the idea that insight (be it spontaneous or instructed) might generally be less deliberate than it is traditionally assumed.

At a more applied level, the present findings underscore the potential of short training interventions as a de-biasing tool (Boissin et al., 2021). Not only might a short training suffice to get people to intuit correctly but even a very cursory or superficial processing of the training explanations when people are reading under time-pressure and are distracted with another task can result in (some) learning. This points to the possible potential in "real-life" applied settings where people might not always have the time, resources, or motivation to deeply reflect on training material or where the training conditions themselves hamper such deep reflection (e.g., from a noisy work-floor or classroom to multitasking during online training).

A possible critique is that our test conditions may have been too challenging, confused people and led them to start guessing after training. Given the typical low pre-training accuracy on the bat-and-ball problem, such random guessing might actually lead to higher accuracies and be erroneously interpreted as a post-training performance boost. However, we controlled for such a guessing confound with our control no-conflict problems in which reasoners show high accuracies throughout. If participants were confused after the training and started responding randomly, they should have shown a clear performance decrease here. In addition, as the individual trial data in Figure 2 shows, participants' performance post-training was highly robust. People who improved, typically did so from the first trial and kept on responding correctly throughout the study. Clearly, any random guessing should have resulted in a more variable pattern.

A related critique is that the learning under our experimental constraints might have been more superficial than in the unrestricted training group. In theory, it is possible that under time-pressure and load burden, participants did not learn the underlying problem logic but simply learned to generally distrust their intuition or developed a simplified solution heuristic (e.g., "it's half of what you think it is"). However, although such a superficial strategy might have helped reasoners on the conflict problems it should have severely hampered their performance on the control no-conflict problems where the intuitively cued response was always correct. The consistent high accuracies on the control problems again argue against such a confound.

At the other end of the spectrum, one might always argue that our test conditions were not challenging enough and did not prevent all possible deliberation. Here it should be noted that our design was carefully piloted. We opted for stringent deadlines and our manipulation checks indicated that in our restricted training groups people took less than half the time they needed when they were

allowed to freely deliberate. In addition, taking more time during the restricted training was not associated with a higher post-training performance. Hence, this argues against the suggestion that slow residual deliberation is driving the training effect. However, at the same time it should be noted that there is no clear, universal threshold (i.e., longer than x seconds or less than x amount of load implies deliberation) that allows us to universally demarcate intuition and deliberation (Bago & De Neys, 2019; De Neys, 2021). Hence, we can never exclude that the allotted time and load allowed some "fast deliberators" to engage in some minimal deliberation. Note that such a position cannot be falsified. We have no empirical criterion to differentiate "fast deliberation" from "pure intuiting". At the same time, this indicates that the label "intuitive insight" needs to be interpreted with some caution. The point here is that the data show that learning can take place under conditions that severely minimize deliberation.

To avoid confusion, it is important to keep some limitations of the present study in mind. It is important to highlight that even when participants were allowed to deliberate, not everyone benefitted from the training. Under restricted deliberation, the number of individuals who remained biased further increased. This indicates that the de-bias training is not perfect. Note that such individual "trainability" differences might be linked to individual differences in "mindware" instantiation (Stanovich, 2018) or the degree to which an individual has managed to automatize the necessary logical operations. Individuals who have been less exposed to and familiarized with the underlying logico-mathematical operations in their school curriculum, might face a harder time to benefit from the current minimalistic interventions. It cannot be excluded that these reasoners do have a more serious knowledge gap and would require a different, more extensive type of de-bias training in which they are taught the underlying principles. This hypothesis is speculative, but the point is that the current de-biasing approach might not work or suffice for everyone.

Relatedly, care should be taken to refrain from overgeneralizing the current findings. The findings do not imply that people can learn any problem solution effortlessly. The critical boundary condition is that we focus on classic bias task which entail rather elementary logico-mathematical principles. As we noted, minimal deliberation presumably suffices here precisely because (most) people have prior knowledge about these principles. Hence, one should refrain from portraying intuition as a magical panacea to arrive at problem insight. Bluntly put, if people generally managed to understand problem solutions intuitively, there would be little point in having them go through years of formal education and explicit, repeated instruction. The current findings need to be interpreted within the proper boundary conditions and do obviously not imply that people should not deliberate in a learning context.

At the same time, we also believe it is important to avoid downplaying the results. Even when the conclusion is restricted to "familiar," elementary principles such as required in the bat-and-ball and

related prototypical bias problems, it is important to establish that these are easily trainable. Not respecting these simple mathematical principles can lead to massive problems in daily life (e.g., think about neglecting the role of base rates when evaluating vaccination efficiency, e.g., Devis, 2021). If we get people to properly apply basic logico-mathematical operations, this may save them from considerable potential harm in their personal and professional lives. The silver lining in the present results is that correcting people's logical reasoning might be more straightforward than often assumed (Milkman et al., 2009). At the very least these findings should motivate theorists and practitioners to take notice and start exploring the role of intuition and deliberation in the generation of problem insight more closely.

## Declaration of competing interest

None.

## Acknowledgments

This study was supported by the Idex Université Paris Cité ANR-18-IDEX-0001 and by a research grant (DIAGNOR, ANR-16-CE28-0010-01) from the Agence Nationale de la Recherche, France.

## Open data statement

Raw data can be downloaded from our OSF page.
(https://osf.io/3b4jy/?view_only=a388443c8fc34310b9f908fe847f077b).

## References

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

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

Bourgeois-Gironde, S., & Van Der Henst, J. B. (2009). How to open the door to System 2: Debiasing the bat-and-ball problem.

Chein, J. M., & Weisberg, R. W. (2014). Working memory and insight in verbal problems: Analysis of compound remote associates. *Memory and Cognition*, 42(1), 67–83. https://doi.org/10.3758/s13421013-0343-4

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. (2012). Bias and Conflict : A Case for Logical Intuitions. *Perspectives on Psychological Science*, *7*(1), 28-38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2017). Bias, Conflict, and Fast Logic. In W. De Neys (Éd.), *Dual Process Theory 2.0* (1re éd., p. 47-65). Routledge. https://doi.org/10.4324/9781315204550-4

De Neys, W. (2021). On dual-and single-process models of thinking. *Perspectives on psychological science*, *16*(6). https://doi.org/10.1177/1745691620964172

De Neys, W., & Bonnefon, J.-F. (2013). The 'whys' and 'whens' of individual differences in thinking biases. *Trends in Cognitive Sciences*, *17*(4), 172-178. https://doi.org/10.1016/j.tics.2013.02.001

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow : Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, *28*(5), 503-509. https://doi.org/10.1177/0963721419855658

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

Devis, D. (2021). Why are there so many vaccinated people in hospital?. Retrieved from https://cosmosmagazine.com/health/covid/why-are-there-so-many-vaccinated-people-in-hospital/

Ellis, J. J., Glaholt, M. G., & Reingold, E. M. (2011). Eye movements reveal solution knowledge prior to insight. *Consciousness and cognition*, *20*(3), 768-776. https://doi.org/10.1016/j.concog.2010.12.007

Evans, J. St. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255-278. https://doi.org/10.1146/annurev.psych.59.103006.093629

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

Evans, J. St. B. T., & 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

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

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

Ghasemi, O., Handley, S., Howarth, S., Newman, I. R., & Thompson, V. A. (2022). Logical intuition is not really about logic. *Journal of Experimental Psychology: General, 151(9)*, 2009-2028. https://doi.org/10.1037/xge0001179

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

Hoover, J., & Healy, A. (2019). The Bat-and-Ball Problem : Stronger Evidence in Support of a Conscious Error Process. *Decision*, *6*(4), 369-380. https://doi.org/10.1037/dec0000107

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

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1 : Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56-64. https://doi.org/10.1016/j.actpsy.2015.12.008

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Strauss, Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited : Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Éds.), *Heuristics and Biases* (1re éd., p. 49-81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

Lawrence, M. A., & Lawrence, M. M. A. (2016). Package 'ez'. R package version, 4(0).

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

Markovits, H., de Chantal, P.-L., Brisson, J., & Gagnon-St-Pierre, E. (2019). The development of fast and slow inferential responding: Evidence for a parallel development of rule-based and belief-based intuitions. *Memory & cognition, 47 (6)*, 1188–1200. https://doi:10.3758/s13421-019-00927-3

Mercier, H., & Claidière, N. (2022). Does discussion make crowds any wiser?. *Cognition*, *222*. https://doi.org/10.1016/j.cognition.2021.104912

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57-74. https://doi.org/10.1017/S0140525X10000968

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How Can Decision Making Be Improved? *Perspectives on Psychological Science*, *4*(4), 379-383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

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

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

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

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

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

Raoelison, M., Boissin, E., Borst, G., & Neys, W. D. (2021b). From slow to fast logic : The development of logical intuitions. *Thinking & Reasoning*, *0*(0), 1-25. https://doi.org/10.1080/13546783.2021.1885488

Raoelison, M., Keime, M., & De Neys, W. (2021a). Think slow, then fast : Does repeated deliberation boost correct intuitive responding? *Memory & Cognition*. https://doi.org/10.3758/s13421-021-01140-x

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin*, *119*(1), 3. https://doi.org/10.1037/0033-2909.119.1.3

Stagnaro, M., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision Making*, *13*, 260-267. https://dx.doi.org/10.2139/ssrn.3115809

Stanovich, K. (2011). *Rationality and the Reflective Mind*. Oxford University Press.

Stanovich, K. (2018). Miserliness in human cognition : The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423-444. https://doi.org/10.1080/13546783.2018.1459314

Stupple, E. J., Pitchford, M., Ball, L. J., Hunt, T. E., & Steel, R. (2017). Slower is not always better: Response-time evidence clarifies the limited role of miserly information processing in the Cognitive Reflection Test. *PloS one*, *12*(11), e0186404. https://doi.org/10.1371/journal.pone.0186404

Stuyck, H., Aben, B., Cleeremans, A., & Van den Bussche, E. (2021). The Aha! moment: Is insight a different form of problem solving?. *Consciousness and cognition*, *90*, 103055. https://doi.org/10.1016/j.concog.2020.103055

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. S. B. T. (2018). Do smart people have better intuitions? *Journal of Experimental Psychology: General, 147*(7), 945–961. https://doi.org/10.1037/xge0000457

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

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147-168. https://doi.org/10.1080/13546783.2013.844729

Topolinski, S., & Reber, R. (2010). Gaining insight into the “Aha” experience. *Current Directions in Psychological Science*, *19*(6), 402-405. https://doi.org/10.1177/0963721410388803

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

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

Wickham, H. (2009) *ggplot2: elegant graphics for data analysis*. Springer New York.

Wickham, H., & Wickham, M. H. (2017). Package 'tidyr'. Easily Tidy Data with 'spread' and 'gather ()' Functions.

# Supplementary material

## A – Material

**Study 1 conflict bat-and-ball items**

- 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 forest one can find 360 wolves and bears in total. There are 300 more wolves than bears. How many bears are there?
- 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 store one can choose between 320 tomatoes and avocados. There are 300 more tomatoes than avocados. How many avocados are there?
- A store manager has bought 310 bananas and kiwis in total. There are 300 more bananas than kiwis. How many kiwis are there?
- A retail clerk has to sort 290 oranges and lemons in total. There are 200 more oranges than lemons. How many lemons are there?
- To make yogurt, a cook has bought 270 apricots and pears. There are 200 more apricots than pears. How many pears are there?
- 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 restaurant clients have been using 250 forks and napkins. There are 200 more forks than napkins. How many napkins are there?
- The kitchen in a restaurant has 240 plates and pans in total. There are 200 more plates than pans. How many pans are there?
- For a convention organizers have bought 230 glasses and cups. There are 200 more glasses than cups. How many cups are there?
- A store is advertising 220 coffee makers and toasters. There are 200 more coffee makers than toasters. How many toasters are there?
- A music store has 210 saxophones and flutes in total. There are 200 more

saxophones than flutes. How many flutes are there?

- A store is showcasing 190 pianos and xylophones in total. There are 100 more pianos than xylophones. How many xylophones are there?

- A music school is renting 170 guitars and harps in total. There are 100 more guitars than harps. How many harps are there?

- 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 park there are 140 adults and children in total. There are 100 more adults than children. How many children are there?

- In a school there are 130 boys and girls in total. There are 100 more boys than girls. How many girls are there?

- A city is employing 120 policemen and firefighters in total. There are 100 more policemen than firefighters. How many firefighters are there?

- 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?

- In a forest there are 640 oak trees and maple trees. There are 600 more oak trees than maple trees. How many maple trees are there?

- In a greenhouse there are 620 dandelions and water lilies. There are 600 more dandelions than water lilies. How many water lilies are there?

- Around a lake there are 610 daisies and jasmine flowers. There are 600 more daisies than jasmine flowers. How many jasmine flowers are there?

- A science fair has gathered 590 inventors and engineers. There are 500 more inventors than engineers. How many engineers are there?

- A scientific committee oversees 580 biologists and mathematicians. There are 500 more biologists than mathematicians. How many mathematicians are there?

- At a convention there are 560 neuroscientists and botanists. There are 500 more neuroscientists than botanists. How many botanists are there?

- In a stadium there are 540 volleyball and baseball players. There are 500 more volleyball than baseball players. How many baseball players are there?

- For a sports event, organizers have invited 530 players and coaches. There are 500 more players than coaches. How many coaches are there?

- A competition features 490 rugby players and runners. There are 400 more rugby players than runners. How many runners are there?
- In a store there are 480 nails and hammers in total. There are 400 more nails than hammers. How many hammers are there?

- On the shelves one can find 470 screws and screwdrivers. There are 400 more screws than screwdrivers. How many screwdrivers are there?

- A city has acquired 430 buses and trains in total. There are 400 more buses than trains. How many trains are there?

**Study 1 conflict bat-and-two-ball items**

- A hat and two ribbons cost $4.20 in total. The hat costs $4.00 more than the two ribbons. How much does one ribbon cost?

- A book and two bookmarks cost $3.60 in total. The book costs $3.00 more than the two bookmarks. How much does one bookmark cost?

- A cheese and two breads cost $2.80 in total. The cheese costs $2 more than the two breads. How much does one bread cost?

- A lime and two oranges cost $4.60 in total. The lime costs $4 more than the two oranges. How much does one orange cost?

- A coffee and two cookies cost $3.80 in total. The coffee costs $3.00 more than the two cookies. How much does one cookie cost?

- A sandwich and two sodas cost $2.40 in total. The sandwich costs $2.00 more than the two sodas. How much does one soda cost?

- A lamp and two pillows cost $3.40 in total. The lamp costs $3.00 more than the two pillows. How much does one pillow cost?

- A necklace and two rings cost $2.20 in total. The lamp costs $3.00 more than the two pillows. How much does one pillow cost ?

**Study 1 intervention bat-and -ball items**

- A hat and a ribbon cost $4.20. The hat costs $4.00 more than the ribbon. How much does the ribbon cost?

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

- 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?

**Study 2 conflict items**

- 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?

- In a jungle there are 340 crocodiles and snakes in total. There are 300 more crocodiles than snakes. How many snakes are there?

- To make juice, a producer used 280 clementines and limes. There are 200 more clementines than limes. How many limes are there?

- An orchestra has bought 180 trombones and clarinets. There are

100 more trombones than clarinets. How many clarinets are there?

- A music company has manufactured 160 banjos and ukuleles. There are 100 more banjos than ukuleles. How many ukuleles are there?

- A university department lists 110 psychologists and statisticians. There are 100 more psychologists than statisticians. How many statisticians are there?
- On a mountain there are 630 pine and eucalyptus trees. There are 600 more pine than eucalyptus trees. How many eucalyptus trees are there?

- NASA has hired 570 astronomers and astronauts in total. There are 500 more astronomers than astronauts. How many astronauts are there?

- The science department has hired 550 physicists and chemists. The science department has hired 550 physicists and chemists.

- 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 event has gathered 520 soccer players and referees. A sports event has gathered 520 soccer players and referees.

- On the shelves one can find 470 screws and screwdrivers. There are 400 more screws than screwdrivers. How many screwdrivers are there?

- A woodwork company has bought 460 drills and hacksaws. There are 400 more drills than hacksaws. How many hacksaws are there?

- In a giveaway there are 450 pliers and scissors in total. In a giveaway there are 450 pliers and scissors in total.

- In a large box there are 440 nuts and bolts in total. There are 400 more nuts than bolts. How many bolts are there?

- An army division has 420 planes and boats in total. There are 400 more planes than boats. How many boats are there?

## B – Equation selection results

After the training, for each participant we presented the bat-and-ball problem and two equations from which the participant had to select the one that needs to be applied to solve the problem. One of the equations corresponds to the correct one, while the other corresponds to the one supposedly applied incorrectly by reasoners. The rationale behind this question is based on the idea that the effect of restricting deliberation during the intervention could affect the level of understanding and learning of the explanations. Thus, if a participant did not understand which logical principle has to be applied to correctly answer the bat-and-ball-type problems, then the participant should be unable to select the equation needed to solve the problem. Thus, we wanted to check whether the participants who were restricted to deliberate during the intervention were able to target the correct equation to be used.

First, Figure S1 shows that in the absence of training, participants performed worse than those who were given the bat-and-ball-problem explanations, regardless of whether the intervention was presented with or without restrictions (all groups comparison: $F(3,161) = 10.86$, $p < .001$, $\eta^2g = .17$, Post-hoc comparison with Holm correction: Control vs Free, $t(82) = 5.62$, $p < .001$, $d = 1.22$; Control vs Fast, $t(82) = 4.69$, $p < .001$, $d = 1.02$; Control vs Fast-load, $t(81) = 2.47$, $p = .05$, $d = .54$). This highlights that when participants are explained how to solve the bat-and-ball problems, they are able to retrieve the logical information behind the correct solution. Also, Figure S1 shows a gradual decrease in the performance to select the correct equation with less deliberation. However, a further analysis that was restricted to the group of improved reasoners (following our individual level classification) showed that this was not the case among those reasoners who actually benefitted from the intervention and responded correctly after training. If anything, improved participants in the Free training group even tended to perform slightly worse than those in the restricted deliberation groups (Free: M = 66.7% , SEM = 11.4, Fast: M = 100%, SEM = 0.0, Fast-load: M = 88.9%, SEM = 11.1, ANOVA on Group factor, $F(2,37) = 3.35$, $p = .05$, $\eta^2g = .15$). In other words, once participants understood the explanations, regardless of whether they were able to deliberate, and responded correctly after the intervention, they were also able to identify the equation leading to the correct answer. However, these equation solving data were collected for exploratory purposes and should be interpreted with caution.

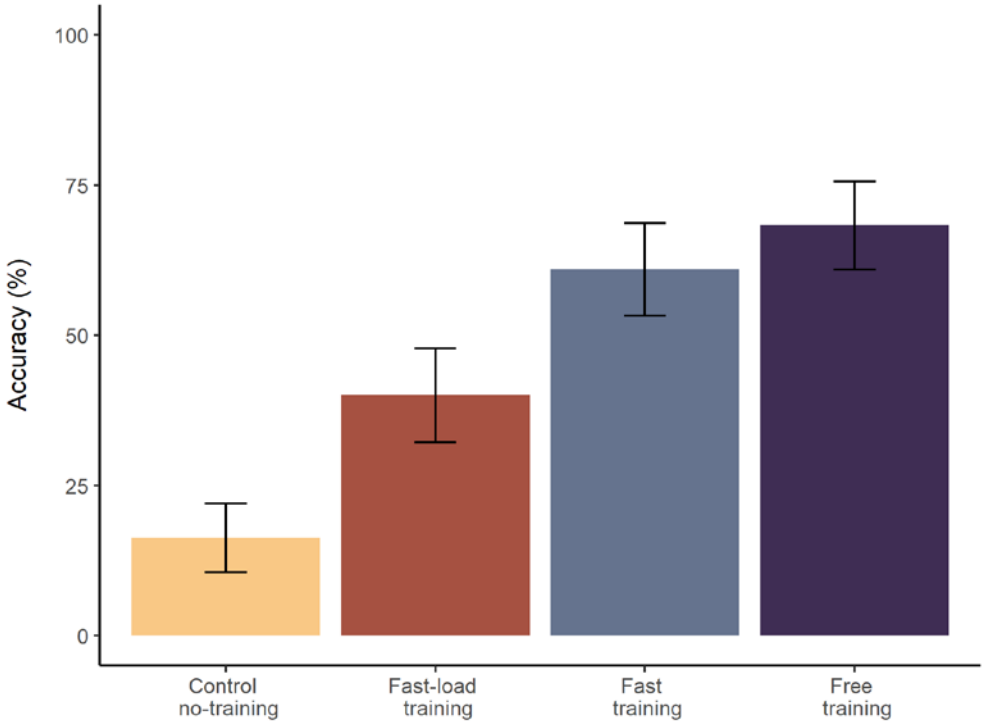


**Figure S1.** Accuracy on equation-selection problem for each intervention group. Error bars represent standard error of the mean (SEM).

## C – Bat-and-two-balls accuracy

In the Fast-load condition, participants were asked to respond to bat-and-two-balls problems before and after the intervention to test whether they were more likely to apply a blind 'halving heuristic' strategy (which should lead to an incorrect answer) after the intervention. Thus, for each participant in the Fast-load group, we calculated the average proportion of correct fast and slow responses for the bat-and-two-balls problems, in each of the two blocks (pre- and post-intervention).

Figure S2 suggests that if participants in the Fast-load group failed to correctly respond to the bat-and-two-balls items before the intervention, they succeed after the intervention. This was particularly the case for the fast trials, $t(39) = 3.57$, $p < .001$, $d = 0.56$. The mean comparison on slow trials failed to reach significance, $t(39) = 1.06$, $p = .30$, $d = 0.17$. In addition, participants who benefited from the very restrictive training, namely the "improved" participants, showed higher performance after the intervention for fast trials (from 0 to 44.4%, $t(9) = 2.87$, $p = .02$, $d = 0.96$) and to a lesser extent for slow trials (from 0 to 16.7%, $t(9) = 2.00$, $p = .09$, $d = 0.67$).

Overall, this suggests that even when deliberation is minimized during the intervention, participants still manage to grasp the logic behind the bat-and-ball problem and do not simply apply a "halving"

heuristic. However, note again that these data were collected for exploratory purposes and should be interpreted with caution.

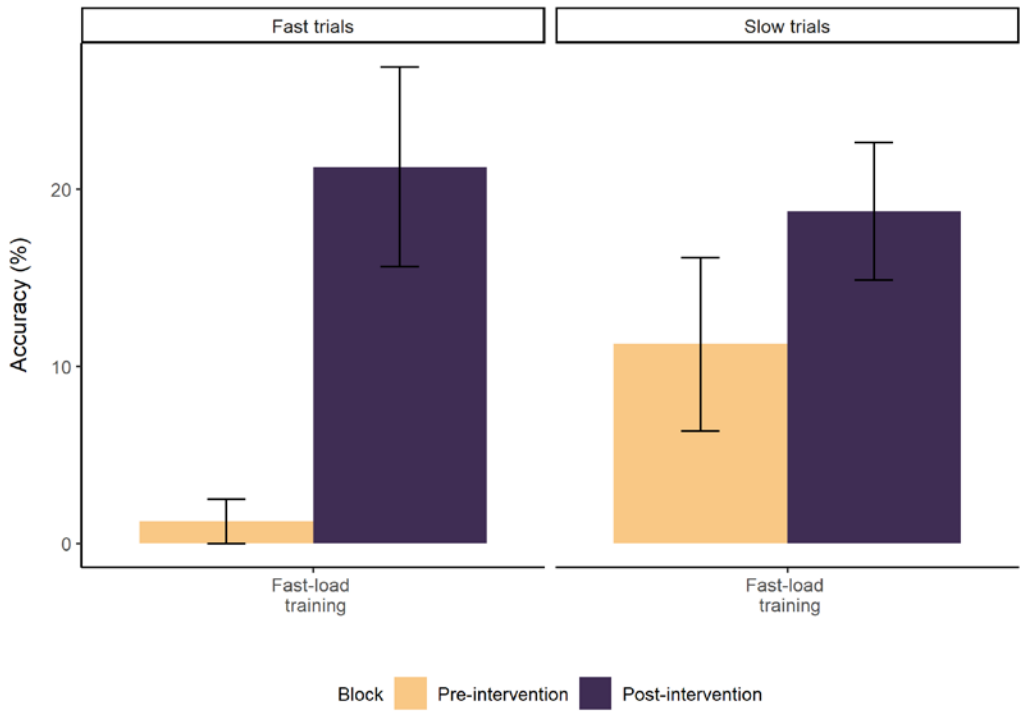


**Figure S2.** Accuracy on bat-and-two-balls problems before and after the Fast-load intervention. Error bars represent standard error of the mean (SEM).

## D – No-conflict problems accuracy

**Table S1.**

Average accuracy (%) for the no-conflict problems (SEM) in Study 1.

| Groups | Fast trial | | Slow trial | |
|---|---|---|---|---|
| | *Pre-intervention* | *Post-intervention* | *Pre-intervention* | *Post-intervention* |
| Control | 91.9 (2.6) | 97.5 (1.5) | 97.3 (1.9) | 96.6 (1.3) |
| Fast-load | 96.3 (2.6) | 86.7 (4.2) | 98.3 (1.0) | 97.3 (1.4) |
| Fast | 95.5 (2.7) | 80.5 (5.2) | 98.8 (0.9) | 92.1 (3.8) |
| Free | 93.7 (2.8) | 79.3 (5.5) | 96.1 (2.3) | 89.6 (1.2) |

**Table S2.**

Average accuracy (%) for the no-conflict problems (SEM) in Study 2 (conjunction fallacy problems).

| Groups | Fast trial | Slow trial |
|---|---|---|
| Fast-load | 94.0 (2.4) | 94.8 (3.8) |
| Fast | 95.2 (2.1) | 95.2 (3.6) |
| Free | 96.1 (1.9) | 90.8 (4.7) |