Cognition 211 (2021) 104645

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

From bias to sound intuiting: Boosting correct intuitive reasoning

Esther Boissin [a,*], Serge Caparos [b,c], Matthieu Raoelison [a], Wim De Neys [a]

[a] Université de Paris, LaPsyDÉ, CNRS, F-75005 Paris, France
[b] Université Paris 8, DysCo lab, Saint-Denis, France
[c] Institut Universitaire de France, Paris, France

ARTICLE INFO

Keywords:
Reasoning
Decision-making
Dual process theory
Heuristics & Biases
De-biasing
Intuition

ABSTRACT

Although human thinking is often biased by erroneous intuitions, recent de-bias studies suggest that people's performance can be boosted by short training interventions, where the correct answers to reasoning problems are explained. However, the nature of this training effect remains unclear. Does training help participants correct erroneous intuitions through deliberation? Or does it help them develop correct intuitions? We addressed this issue in three studies, by focusing on the well-known Bat-and-Ball problem. We used a two-response paradigm in which participants first gave an initial intuitive response, under time pressure and cognitive load, and then gave a final response after deliberation. Studies 1 and 2 showed that not only did training boost performance, it did so as early as the intuitive stage. After training, most participants solved the problems correctly from the outset and no longer needed to correct an initial incorrect answer through deliberation. Study 3 indicated that this sound intuiting sustained over at least two months. The findings confirm that a short training can boost sound reasoning at an intuitive stage. We discuss key theoretical and applied implications.

1. Introduction

Decades of research have shown that human reasoning and decision making are sometimes biased by intuition-related heuristics. People tend to base their judgments on quick and intuitive impressions rather than on more costly deliberative thinking (e.g., Evans, 2008; Kahneman, 2011; Stanovich & West, 2000; Thompson, Prowse Turner, & Pennycook, 2011). While those intuitions can sometimes be useful, they can also conflict with basic logical, probabilistic and mathematical considerations (Evans, 2008; Kahneman & Frederick, 2005). One of the problems that 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, in the bat-and-ball problem, our reasoning is biased despite the problem solution being based on a simple algebraic equation: "X+Y=1.10, Y=1+X, Solve for X", which any adult who has received formal education encountered in secondary school mathematics (Hoover & Healy, 2017). More interestingly, the "10 cents" answer is given in a majority of cases (Toplak, West, & Stanovich, 2014; Travers, Rolison, & Feeney, 2016), even in samples composed of highly qualified university students (Bourgeois-Gironde and Van Der Henst, 2009; Frederick, 2005), and even after repeated exposure to the problem (Raoelison & De Neys, 2019; Stagnaro, Pennycook, & Rand, 2018). Although the "5 cents" answer does not require complex mathematical operations, it is not directly accessible to most reasoners.

One explanation for this phenomenon relies on the idea that human reasoning arises from the interaction between two types of processes or "systems": the intuitive system and the deliberative system (e.g., Epstein, 1994; Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996). According to this dual-process model, human reasoning is biased because reasoners tend to make extensive use of the intuitive, fast and inexpensive system, at the expense of the deliberative system, which is slow and demanding in terms of cognitive resources. Reasoners who manage to solve the problem correctly would correct their initially-generated intuitive response (e.g., the "10 cents" answer) after completing deliberative calculations (Evans & Stanovich, 2013; Kahneman, 2011; Kahneman & Frederick, 2005; Morewedge & Kahneman, 2010). Because most reasoners tend to minimize demanding

* Corresponding author at: LaPsyDE (CNRS & Université de Paris) Sorbonne - Labo A. Binet 46, rue Saint Jacques, 75005 Paris, France.
E-mail address: boissinesther@gmail.com (E. Boissin).

https://doi.org/10.1016/j.cognition.2021.104645
Received 21 October 2020; Received in revised form 19 January 2021; Accepted 19 February 2021
Available online 3 March 2021

computations (Kahneman, 2011), they would apply the intuitive system by default and simply stick to the answer that quickly comes to mind without considering that the correct answer could be different from the intuitively-generated one.

However, although the viewpoint of a deliberative corrective system has long dominated the field, some recent studies have shown that correct responses can sometimes be intuitive and do not necessarily need to be corrected (e.g., Bago & De Neys, 2017, 2019; Newman, Gibb, & Thompson, 2017). These studies adopted a two-response paradigm (Thompson et al., 2011) in which participants were asked to provide two consecutive answers to a given problem. In order to prevent the involvement of the deliberative system for the initial answer, participants had to provide an intuitive response under time-pressure and, at the same time, perform a secondary memory-task that is supposed to burden cognitive resources (Bago & De Neys, 2019). Immediately afterwards, they could take all the time they needed to think about the problem before giving a final answer. Results showed that sound reasoners often already give a correct answer at the initial (intuitive) stage (Bago & De Neys, 2017, 2019; Newman et al., 2017; Raoelison & De Neys, 2019; Thompson et al., 2011). Importantly, reasoners produced more final correct answers for which the initial answer was also correct, than final correct answers for which the initial answer was incorrect (Bago & De Neys, 2017, 2019; Janssen, Raoelison, & de Neys, 2020; Raoelison, Keime, and De Neys, 2021; Raoelison, Thompson, and De Neys, 2020; Raoelison & De Neys, 2019). Those results suggest that, sound reasoners do not necessarily need to deliberate to correct their "erroneous" intuitions, since intuitions actually lead to correct responses. Applied to the bat-and-ball problem, the two-response paradigm highlights that some reasoners can automatically use basic logico-mathematical principles without necessarily engaging the deliberative system and its corrective function (Bago & De Neys, 2019). However, even though correct answers *can* be generated intuitively, they are overall still rare (Bago & De Neys, 2017, 2019; Janssen et al., 2020; Newman et al., 2017; Raoelison & De Neys, 2019; Thompson et al., 2011). That is, most reasoners remain "biased" and fail to respond correctly. In this study, we investigate whether we can boost correct intuitive responses with a short training intervention.

Recent de-biasing studies have shown that a short explanation about the notorious bat-and-ball problem helps reasoners produce a correct response (Claidière, Trouche, & Mercier, 2017; Hoover & Healy, 2017; Morewedge et al., 2015; Purcell, Wastell, & Sweller, 2020; Trouche, Sander, & Mercier, 2014). Once the problem has been explained to reasoners, they manage to solve structurally similar problems afterwards. However, no study has explored the nature of the training effect: Are participants after the training better able to deliberate and correct an "erroneous" intuitive response, or does the training help participants to intuit the correct solution (i.e., after training correct responding no longer requires a corrective deliberation process)?

Clearly, if a de-biasing training actually helps people intuit correctly, this would have great potential. Although it can be laudable to help people to deliberate more, in many daily life situations they will simply not have the time (or resources/motivation) to deliberate. Hence, if de-biasing interventions only help people to deliberately correct erroneous intuitions, their impact may be suboptimal. The potential benefits of training sound intuiting are rife in this respect.

Interestingly, indirect evidence lends some credence to the "trained intuitor" point of view. For example, it has been shown that repeated exposure to bat-and-ball problems, with no explanation given about the correct solution, sometimes leads to spontaneous insight. Some participants are biased at first but after a couple of trials do start to answer correctly (Raoelison & De Neys, 2019). Two-response findings indicate that after such learning occurs, the intuitive responses on the later trials are typically correct too. Although this spontaneous learning occurs only for a handful of reasoners, it seems that, people can easily switch from incorrect to correct intuitive responding once they grasp how to solve the problem (Raoelison & De Neys, 2019). Thus, if a training intervention could generate insight about the solution strategy, then it may be that the same training could boost correct intuitive responses. Just like natural sound reasoners, we may be able to lead biased reasoners, through a simple training intervention, to intuitively generate correct answers.

In the present work, we conducted three studies in which we explored the impact of a training intervention on participants' reasoning performance, using the bat-and-ball problem. In all three studies, we contrasted participants' reasoning performance before and after a short training session and compared their performance to that of participants who received no training (the control group). We measured performance using a two-response paradigm (Thompson et al., 2011) in order to determine whether the intervention affected participants' intuitive and/or deliberative reasoning. The structure of the experiment was the same in all three studies: Participants always performed two blocks of problems (pre-intervention and post-intervention) which were separated by an intervention block, where participants were given an explanation about the bat-and-ball problem (training group) or no explanation (control group).

Before running our three main studies we ran a pre-test study (as a manipulation check), to ensure that we could train participants to solve the bat-and-ball problem with our intervention. In Study 1 we then tested the nature of the training by using a two-response paradigm. Study 2 tested whether we could replicate our findings with an improved design. Study 3 re-tested the participants from Study 2 two months later to explore whether the training effect sustained over time.

2. Pre-test study

The purpose of the pre-test study was to evaluate the efficiency of our training procedure, which consisted of two short explanations describing the strategy that should be used to solve bat-and-ball problems. We presented three problems to the participants, always in that order: (1) First, the original bat-and-ball problem, used to measure participants' basic performance in the absence of an explanation, (2) second, a structurally similar version of the bat-and-ball problem (with different surface content), which was preceded by a short explanation about how to solve this type of problem, and which allowed us to measure the effect of an explanation on performance, and finally, (3) a third bat-and-ball problem, presented after a second explanation, and for which participants only had 6.5 s to provide their answer. This last problem was added for exploratory purposes. Although the pre-test did not adopt a proper two-response design, the trial could give us a rough indication of whether the given explanation can affect participants' intuitive performance (i.e., when the possibility to deliberate is reduced). Note that the data of this pre-test study were collected just after collection of the data already presented in Raoelison, Keime, and De Neys (2021), using the same participants.

2.1. Methods

2.1.1. Participants

One hundred and twenty-three participants (79 females, Mean age = 34.9 years, SD = 12.9 years[1]) were recruited online using the Prolific Academic website (http://www.prolific.ac). In order to take part, participants had to be native English speakers from Canada, Australia, New Zealand, the USA, or the UK. Among them, two participants did not complete secondary school, 48 participants reported secondary school as their highest level of education, and 73 reported a university degree. We compensated participants for their time at the rate of £5 per hour.

Note that as part of our procedure (see below) we asked participants whether they were familiar with the original bat-and-ball problem. In total, 19 participants reported having come across the problem before

[1] Due to a technical error, the age of three participants was missing.

and also provided the correct "5 cents" response. We excluded them 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 thus kept the remaining 104 participants in the analyses.

2.1.2. Materials & procedure

First, participants were shown the original bat-and-ball problem taken from Frederick (2005):

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

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. This first problem was used to obtain a performance baseline. After participants had provided their response, they saw a short explanation about how to solve the bat-and-ball problem, which read:

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

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

However, the problem said they cost $1.10 together.

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

The explanation was adapted from previous studies (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, we 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.

They were then presented with a second version of the bat-and-ball problem, which shared the same structure as the standard problem but had a different superficial content (Bago, Raoelison, & De Neys, 2019):

A banana and an apple cost $1.40.

The banana costs $1.00 more than the apple. How much does the apple cost?

Again, response time was unlimited, allowing participants to deliberate before answering. After they provided their answer, an explanation was presented using the same principle as mentioned previously but adapted to match the content of the second problem.

Finally, participants saw a third problem, taken from Raoelison and De Neys (2019). Unlike the first two problems, this third problem was accompanied by four response choices: (1) the correct response (i.e., which would be "5 cents" in the original bat-and-ball), (2) the intuitively cued "heuristic" response (i.e., "10 cents" in the original bat-and-ball), (3) a foil option which was the sum of correct and heuristic answers (i.e., "15 cents"), and (4) a second foil option which was the second greatest common divider (i.e., "1 cent"). Mathematically speaking, the correct equation to solve the standard bat-and-ball problem is: "$1.00 + 2x = $1.10", instead, people are thought to be intuitively using the "$1.00 + x = $1.10" equation to determine their response (Kahneman, 2011). The latter equation was used to determine the "heuristic" answer option, and the former to determine the correct answer option for this problem. The four response choices appeared in a random order. For instance:

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

There are 100 more pens than pencils. How many pencils are there?

o *25*
o *50*
o *75*
o *10*

A second difference between the third and the first two problems was that there was a limited time to answer. The response time deadline was based on previous studies and was assumed to minimize deliberation (Bago & De Neys, 2019; Raoelison & De Neys, 2019; Thompson et al., 2011). Thus, it allowed us to get some indication of the possible effect of the explanation on one's more "intuitive" performance.

The third problem was presented using the following procedure: A fixation cross was first shown for 1000 ms. We then presented the first sentence of the problem (i.e., "In an office there are 150 pens and pencils in total."). After 2000 ms, the question appeared below the first sentence (i.e., "There are 100 more pens than pencils. How many pencils are there?") and both sentences remained on screen for an additional 4000 ms. Finally, the first sentence and the question were replaced by the four response options and participants had a maximum of 2500 ms to select their response. In total, participants had a maximum of 6500 ms to read the question, solve the problem and select their answer. For this last problem, they were explicitly instructed to respond as fast as possible. Note that participants were familiar with the fast-response procedure given that right before the pre-test they had participated in a reasoning study that adopted a similar procedure (data presented in Raoelison, Keime, and De Neys, 2021). After having answered to the three problems, participants filled in their demographic information.

2.1.3. Trial exclusion

For the third problem, the missed trials were discarded, and we analysed the remaining 89 trials (representing 85.6% of all third-problem trials).

2.1.4. Statistical analyses

The data were processed and analysed using the R software (R Core Team, 2017) and the following packages (in alphabetical order): *dplyr* (Wickham, Francois, Henry, & Muller, 2020), *ez* (Lawrence, 2016), *ggplot2* (Wickham, 2016), and *tidyr* (Wickham & Henry, 2020).

2.2. Results and discussion

2.2.1. Accuracy

A comparison of the mean response accuracies for the first and second problem showed that participants gave more correct responses to the second problem (M = 68.3%, SE = 4.6) than to the first one (M = 21.2%, SE = 4.0), $Z = 1225.0$, $p < .001$, $r = 0.69$. The short explanation given after the first problem thus boosted participants' performance on the second ('deliberation-allowed') problem. This result replicates the training effect observed in previous studies (Claidière et al., 2017; Hoover & Healy, 2017; Morewedge et al., 2015; Purcell et al., 2020; Trouche et al., 2014). After a short explanation, the majority of reasoners manages to solve the bat-and-ball problem.

We then compared the mean response accuracy for the third (limited-time) problem to that of the second and first problem. Although performance on the third (M = 53.9%, SE = 5.3) problem was slightly lower than that on the second problem, (M = 67.4%, SE = 5.0), it was still more than twice as high as that on the first problem (M = 24.7%, SE = 4.6), $Z = 14.5$, $p < .001$, $r = 0.52$. This last result tentatively suggests that the explanations might have boosted participants' ability to provide correct intuitive responses to bat-and-ball-like problems. That is, once participants understand the underlying logic, they can apply it intuitively and no longer need to deliberate to correct an erroneous intuition.

3. Studies 1 and 2

Studies 1 and 2 present a proper test of our hypothesis concerning the nature of the training effect. In both studies we presented bat-and-ball-like problems using the two-response paradigm (Thompson et al., 2011), in which participants had to give an initial response – under severe time-pressure and cognitive load – followed by a final response – without any constraint (e.g. Bago & De Neys, 2019). Participants performed three blocks of trials, namely, (1) a pre-intervention, (2) an intervention, and (3) a post-intervention block. There were two groups of participants, a training group and a control group. While the training group received explanations about how to solve the bat-and-ball problem, during the second "intervention" block of trials, the control group received no such explanation during the second block of trials.

Study 2 introduced a number of potential design optimizations (i.e., longer blocks and additional "bat-and-two-balls" control trials). Given that the general method and results of Studies 1 and 2 were highly similar we will present them alongside each other. Unique features will be explicitly highlighted.

3.1. Methods

3.1.1. Preregistration

The study design and hypotheses were preregistered on the Open Science Framework (http://osf.io/qx7fc). No specific analyses were preregistered.

3.1.2. 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, 99 individuals participated in Study 1 (63 females and 4 gender-neutral, M = 35.6 years, SD = 13.9; 49 participants randomly assigned to the training group and 50 to the control group), and 99 individuals participated in Study 2 (74 females and 1 neutral-gender, M = 34.6 years, SD = 13.7; 50 participants were randomly assigned to the training group and 49 to the control group). In Study 1, one participant had not completed secondary school, 42 participants had secondary school as their highest level of education, and 54 reported a university degree. In Study 2, five participants reported a level of education lower than secondary school, 42 participants reported secondary school as their highest level of education, and 52 reported a university degree. We compensated participants for their time at the rate of £5 per hour.

We again screened for familiarity with the original bat-and-ball problem (during the intervention, see below). In Study 1, 15 participants reported that they already knew the problem and also provided the correct ("5 cents") response. They were excluded from the analyses (e.g., see Bago & De Neys, 2019) and we kept 84 participants (39 in the training group and 45 in the control group). In Study 2, nine participants reported having seen the bat-and-ball problem before and provided the correct ("5 cents") response. They were excluded, leaving 90 participants in the analyses (47 in the training group and 43 in the control group).

3.1.3. Materials

The studies were composed of three blocks presented in the following order: a pre-intervention, an intervention, and a post-intervention block. In total, each participant had to solve 24 problems in Study 1 and 30 problems in Study 2. In Study 1, participants responded to four conflict, four no-conflict and four transfer problems (two neutral and two CRT-like problems, in that order, see below) during the pre-intervention, and again the same number of problems during the post-intervention. In Study 2, during the pre-intervention, participants responded to four conflict, four no-conflict, four transfer and two "bat-and-two-balls" problems (see further). During the post-intervention, they responded to six conflict, four no-conflict, four transfer and two "bat-and-two-balls" problems. All the problems are presented in the Supplementary Material Section A.

3.1.3.1. Bat-and-ball problems. In both Studies 1 and 2, we presented problems taken from Raoelison and De Neys (2019). They were modified versions of the bat-and-ball problem, which used quantities instead of prices (like the third item in the Pre-test Study; Bago & De Neys, 2019; Janssen et al., 2020; Raoelison & De Neys, 2019). They were presented using a free-response format, where participants typed in their response using a computer keyboard (e.g., see Bago & De Neys, 2019).

Some 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, 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, Rossi, & Houdé, 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, accuracy should be at ceiling (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).

Two sets of problems were used in order to counterbalance problem content: The conflict problems in one set were the no-conflict problems in the other, and vice-versa. The presentation order of the conflict and no-conflict problems was randomized in each set. Participants were randomly assigned to one of the two sets for each block.

3.1.3.2. Transfer problems. In addition to the bat-and-ball problems, we used other types of reasoning problems to test whether the "bat-and-ball" training effect could transfer to untrained problems.

Our main interest here were four Cognitive Reflection Test (CRT)-like items that were presented at the end of the pre-intervention and post-intervention block. As the bat-and-ball problem, these items are designed to cue a strong biasing heuristic response and consequently show also very low accuracy rates (Frederick, 2005). However, they require a different solution strategy than the bat-and-ball problem. Two problems were based on the "race" problem from Thomson and Oppenheimer (2016):

If you are running a race and you pass the person in the second place, what place are you in?

Here, the heuristic response is "first place" and the correct response is "second place".

The other two problems were based on the "widget" problem (Frederick, 2005).

If it takes 4 hours for four carpenters to make 4 chairs
How long would it take for 40 carpenters to make 40 chairs?

Here, the heuristic response is "40 hours" and the correct response is "4 hours".

In addition to the CRT-like problems our study also included four neutral[2] problems taken from Raoelison, Thompson, and De Neys (2020). These neutral problems are basic arithmetic word problems which—unlike the conflict, no-conflict, or CRT-like problems—are not expected to cue a strong heuristic answer. For example:

In a bar there are forks and knives.
There are 20 forks and twice as many knives.
How many forks and knives are there in total?

These relatively simple problems are traditionally used to track people's knowledge of underlying logico-mathematical building blocks or "mindware" (Stanovich, 2011). Critically, however, although solving

[2] Due to a coding error, the last neutral problem featured in the post-intervention was discarded from the analysis in Study 1.

the problems requires using similar basic mathematical operations (i.e., addition, multiplication) they do not feature the exact same substitution equation as the bat-and-ball problem (e.g., Y = 2X. X = 20. Y + X =? vs X + Y = 220. Y = X + 200. X =?). Hence, we reasoned that these problems could also be used to test for a potential transfer effect. They allowed us to explore whether the training boosted participant's basic arithmetic word problem solving more generally.

3.1.3.3. Bat-and-two-balls problems. In Study 2, we introduced a new type of problem in order 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 control problems should allow us to identify such a blind "halving heuristic" we wanted to build some additional control into Study 2. The following is an example of what we refer to as the "bat-and-two-balls" problem:

A bat and two balls cost $2.60 in total.
The bat costs $2 more than two balls.
How much does one ball cost?

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. But critically, the basic equation substitution logic is completely similar. If you understand the bat-and-ball structure, then in theory you should also manage to solve the bat-and-two-balls problem. In the new bat-and-two-balls problems, we expected three types of responses: Heuristic (x = $2.60 - $2), halving (x = ($2.60 - $2) /2), and correct (x = ($2.60 - $2) / 4). If the training intervention only cues a halving strategy, then the training should increase correct responses only for the standard "bat-and-ball" problems, and not for the new "bat-and-two-balls" problems. However, if the training intervention does help participants grasp the underlying logic of the problems, then the training should increase correct responses for both the standard "bat-and-ball" and the new "bat-and-two-balls" problems.

3.1.3.4. Justification. After the last problem of the post-intervention block, which was always a conflict problem, participants were asked to type in a justification for their final response (see Supplementary Material Section B for full methodological details). Previous work (e.g., Bago & De Neys, 2019) indicated that correct bat-and-ball reasoners typically manage to correctly justify their answer (e.g., "It's 5 cents because a 5 cents ball and $1.05 bat gives total of $1.10"). Given a coding error, the justifications were not accurately recorded in Study 1. However, Study 2 results indicated that the majority of correct responses was indeed correctly justified (training group: 22 correct justification out of 32 correct responses; control group: 4 correct justifications out of 5 correct responses, see Supplementary Material Section B). Note that the justification was untimed and retrospective. It was collected for exploratory purposes and does obviously not allow drawing any conclusions with respect to the intuitive or deliberate nature of participants' processing.

3.1.3.5. Intervention block. During the intervention block of Study 1, the participants tried to solve one standard and one modified (banana-and-apple) version of the bat-and-ball problem. In Study 2, they tried to solve one standard and two modified (banana-and-apple and magazine-and-banana) versions of the bat-and-ball problem.

They had an unlimited time to give their response. For the standard problem only, participants indicated whether they had seen it before. Participants in the training group were given an explanation of the correct solution after haven given their response to each problem (see Pre-test Study). Participants in the control group received no explanation after they responded. We added the extra intervention block problem (+ explanation) in Study 2 because we expected it could further boost the training effect we observed in Study 1.

3.1.3.6. Two-response format. For both the pre- and post-intervention blocks, participants responded to each problem using a two-response procedure, where they first provided a 'fast' answer, directly followed by a second 'slow' answer (Thompson et al., 2011). This method allowed us to capture both an initial "intuitive" response and then a final "deliberate" one. To minimize the possibility that deliberation was involved in producing the initial 'fast' response, participants had to provide their initial answer within a strict time limit while performing a concurrent cognitive load task (see Bago & De Neys, 2017, 2019; Raoelison & De Neys, 2019). The load task was based on the dot memorization task (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001) given that it had been successfully used to burden executive resources during reasoning tasks (e.g., De Neys, 2006; Franssens & De Neys, 2009). Participants had to memorize a complex visual pattern (i.e., 4 crosses in a 3 × 3 grid) presented briefly before each reasoning problem. After their initial (intuitive) response to the problem, participants were shown four different patterns and had to identify the one that they had memorized (see Bago & De Neys, 2019, for more details).

As in Bago and De Neys (2019), a time limit of 8 s was chosen for the initial response, based on pretesting that indicated 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-restrictions indicated that participants typically need over 30s to solve the bat-and-ball problem correctly (Johnson, Tubau, & De Neys, 2016; Stupple, Pitchford, Ball, Hunt, & Steel, 2017). Hence, by all means the 8 s deadline is challenging. In addition, participants are also under secondary task load when giving their initial response. Obviously, the time limit and cognitive load were applied only for the initial response, and not for the final one where participants were allowed to deliberate (see below).

3.1.4. 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, for which they would have to provide two consecutive responses. They were told that we were interested in their very first, initial answer that comes to mind and that – after providing their initial response – they could reflect on the problem and take as much time as they needed to provide a final answer. In order to familiarize themselves with the two-response procedure, they first solved two unrelated practice problems. Next, they familiarised themselves with the cognitive load procedure by solving two load trials and, finally, they solved two problems which included both cognitive load and the two-response procedure.

Fig. 1 shows a typical trial, which started with a fixation cross for 2000 ms, followed by the first sentence of the problem (e.g., "In an office, there are 150 pens and pencils in total.") for 2000 ms, and followed by the visual matrix for the cognitive-load task for 2000 ms. Then the full problem was presented, at which point participants had 8000 ms to give their initial answer. After 6000 ms the background of the screen turned yellow to warn participants that they only had a short amount of time left to answer. If they had not provided an answer before the time limit, they were given a reminder that it was important to provide an

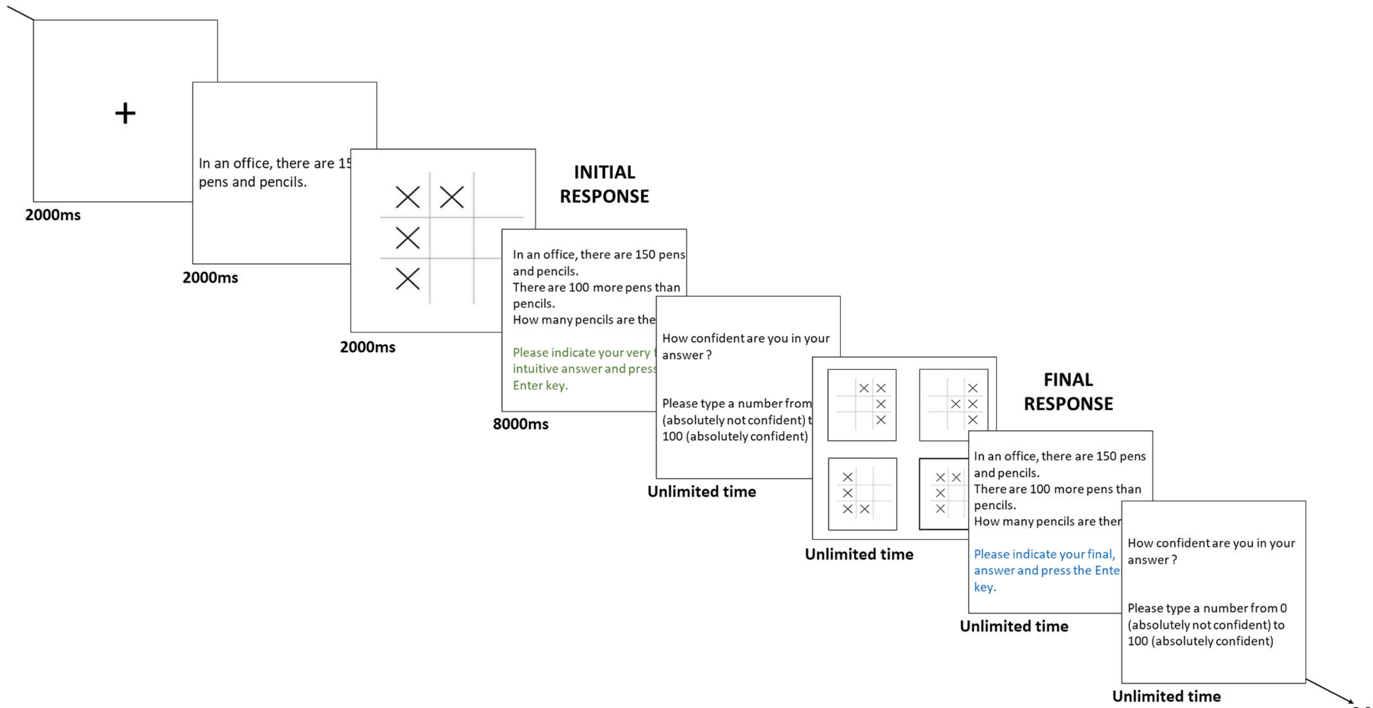

Fig. 1. Time course of a complete two-response trial.

answer within the time limit on subsequent trials. Participants were then asked to enter how confident they were with their response (from 0%, absolutely not confident, to 100%, absolutely confident; note that this confidence rating was not used for CRT-like transfer problems). Then, they were presented with four visual matrices and had to choose the one that they had previously memorized. They received feedback as to whether their memory-response was correct. If the answer was not correct, they were reminded that it was important to perform well on the memory task on subsequent trials. Finally, the same reasoning problem was presented again, and participants were asked to provide a final deliberate answer (with no time limit) and, once again, to indicate their confidence level.

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

3.1.5. *Trial exclusion*

In Study 1 and Study 2, we discarded trials in which participants failed to provide their initial answer before the deadline (5.6% of all Study 1 trials and 3.1% of all Study 2 trials) or failed to pick the correct matrix in the load task (13.4% of the remaining trials in Study 1 and 14.8% of the remaining trials in Study 2), and we analysed the remaining 81.7% of all Study 1 trials and the remaining 82.5% of all Study 2 trials. On average, each participant contributed 19.2 (SD = 3.1) trials out of 24 in Study 1 and 22.4 (SD = 2.7) trials out of 30 in Study 2.

3.2. *Results and discussion*

3.2.1. *Bat-and-ball response accuracy*

For each participant, we calculated the average proportion of correct initial and final responses for the conflict problems, in each of the two blocks (pre- and post-intervention). We analysed the data using mixed-design ANOVAs on initial and final accuracies with Block (pre- vs post-intervention) as a within-subjects factor and Group (training vs control) as a between-subjects factor.

First, we focus on accuracies for the final responses. Fig. 2 shows that most reasoners, from both the control and training group, failed to solve the conflict problems before the intervention (respectively, M = 17.2%, SE = 5.1, and M = 13.8%, SE = 5.6, in Study 1, and M = 6.4%, SE = 3.6, and M = 15.3%, SE = 4.7 in Study 2). The average performance of both groups improved after the intervention, however, the increase in performance was larger in the training group (increase of M = 34.4%, SE = 6.6, in Study 1, and M = 47.2%, SE = 6.0, in Study 2) than in the control group (increase of M = 9.4%, SE = 3.6, in Study 1, and M = 5.7%, SE = 2.8, in Study 2); accordingly, the Block x Group interaction was significant both in Study 1, $F(1,81) = 12.0$, $p < .001$, $\eta^2 g = 0.02$, and in Study 2, $F(1,88) = 32.1$, $p < .001$, $\eta^2 g = 0.09$. In Study 1, the ANOVA also showed that, while the main effect of Block was significant, $F(1,81) = 37.1$, $p < .001$, $\eta^2 g = 0.07$, the main effect of Group was not, $F(1,81) = 1.3$, $p = .26$, $\eta^2 g = 0.013$. In Study 2, both the main effects of Block, $F(1,88) = 52.4$, $p < .001$, $\eta^2 g = 0.13$, and Group, $F(1,88) = 22.9$, $p < .001$, $\eta^2 g = 0.13$, were significant. These results are consistent with previous training studies and indicate that explaining the bat-and-ball led to a substantial improvement in reasoning performance.

To explore whether the training improved people's intuitive reasoning performance, we repeated the analyses on accuracies of the initial responses. The results were fully consistent (see Fig. 2). Once again, most reasoners – from both control and training groups – failed to solve the conflict problems before the intervention (respectively, M = 11.5%, SE = 3.1, and M = 11.2%, SE = 4.8, in Study 1, and M = 5.2%, SE = 3.3 and M = 8.3%, SE = 3.7, in Study 2), but improved after the intervention. However, the improvement was higher in the training group (performance increase of M = 30.0%, SE = 6.6, in Study 1, and M = 45.7%, SE = 6.1, in Study 2) than in the control group (performance increase of M = 11.9%, SE = 4.3, in Study 1, and M = 6.1%, SE = 2.9, in Study 2); accordingly, the Block x Group interaction was again significant both in Study 1, $F(1,81) = 5.6$, $p = .02$, $\eta^2 g = 0.02$, and in Study 2, $F(1,88) = 32.1$, $p < .001$, $\eta^2 g = 0.10$. The ANOVA in Study 1 also showed that, while the main effect of Block was significant, $F(1,81) = 29.8$, $p < .001$, $\eta^2 g = 0.08$, the main effect of Group was not, $F(1,81) = 1.6$, $p = .21$, $\eta^2 g = 0.01$. In Study 2, both main effects of Block ($F(1,88) = 54.6$, $p < .001$, $\eta^2 g = 0.16$) and Group ($F(1,88) = 18.54$, $p < .001$, $\eta^2 g = 0.13$) were significant.

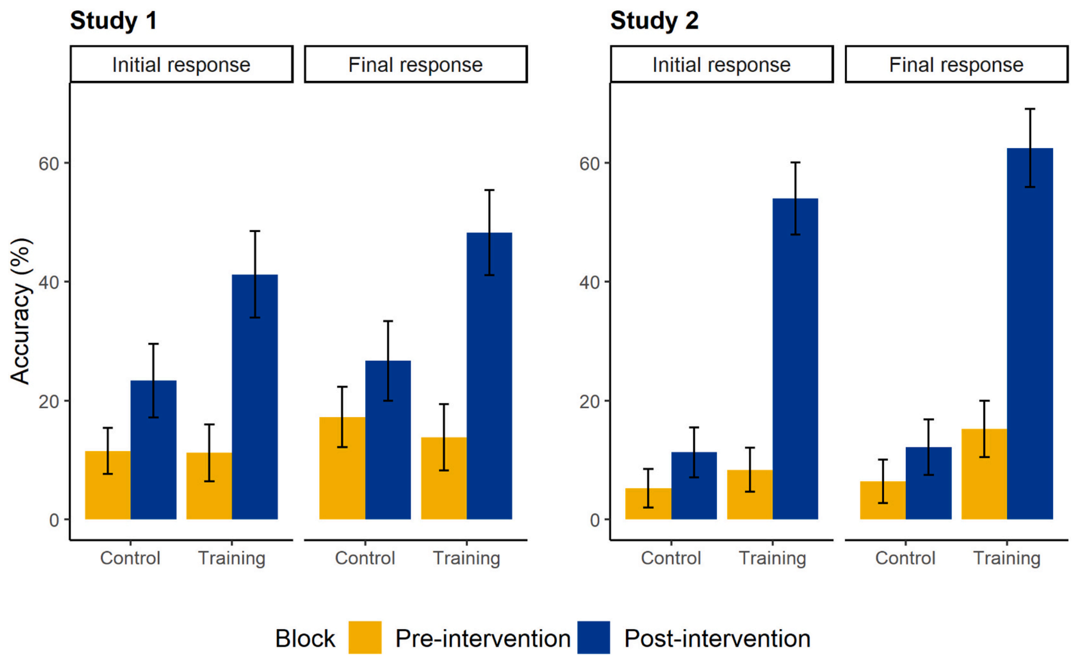

Fig. 2. Average initial and final accuracy on conflict problems in Study 1 and 2. Error bars are standard errors.

In sum, the data showed that the training intervention helped participants to produce more correct responses. Critically, this improvement was shown not only for final "deliberate" responses, for which participants had time and resources to reflect on their response, but also for initial "intuitive" responses, where deliberation was minimized.

For completeness, we also ran a mixed-design ANOVA on accuracies using Block (pre- vs post-intervention) and Response-stage (initial vs. final) as within-subjects factors, and Group (training vs control) as a between-subjects factor, to test whether the intervention effect differed between initial and final responses. The analysis revealed that the interaction between the three factors was not significant, in neither Study 1 nor Study 2, respectively, $F(1,81) = 1.7$, $p = .19$, $\eta^2 g = 0.005$, and $F(1,87) = 0.2$, $p = .70$, $\eta^2 g = 0.00$, showing that the effects of the control and training interventions were similar for initial and for final responses (see Fig. 2).

As expected, for the no-conflict control problems, we observed that performance was at ceiling, with grand means of 94.6% (SE = 1.2) for initial accuracy, and 96.2% (SE = 1.2) for final accuracy in Study 1, and grand means of 93.8% (SE = 1.2) for initial accuracy and 96.3% (SE = 1.0) for final accuracy in Study 2 (See Supplementary Material Section C).

Finally, note that in Study 2 we gave people an additional explanation during the intervention block (i.e., 3 vs 2 problems). We wanted to explore whether this further boosted the training effect we observed in Study 1. A between study comparison indicated that both the initial accuracy increase (30.0% increase in Study 1 vs 45.7% increase in Study 2), and final accuracy increase (34% increase in Study 1 vs 47.2% increase in Study 2) were higher after training in Study 2. Analyses only revealed a trend for the initial accuracy difference increase ($t(83) = 1.73$, $p = .09$) and no significance for the final accuracy difference increase ($t(83) = 1.34$, $p = .18$). Nevertheless, as our analyses showed, the training effect was clearly observed in both studies.

3.2.2. *Direction of change*

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

Fig. 3 plots the direction of change distribution for Studies 1 and 2, for the conflict problems in both the pre- and post-intervention blocks. Fig. 3 shows that, in both studies, before the intervention, participants in the control group were more likely to produce "00" patterns (81.3% and 93.6%, for studies 1 and 2 respectively) than "11" patterns (10.1% and 5.0%) or "01" patterns (7.9% and 1.4%). The same tendency was observed in the training group ("00" patterns: 84.9% and 85.3%, "11" patterns: 11.9% and 6.2%, "01" patterns: 3.2% and 6.8%). These results are in line with several previous studies, which have shown that a majority of participants is biased and fails to solve the bat-and-ball problem, even when allowed to deliberate (Bago & De Neys, 2019; Janssen et al., 2020; Raoelison, Keime, & De Neys, 2021, Raoelison, Thompson, & De Neys, 2020; Raoelison & De Neys, 2019).

After the intervention, similar results were observed for participants in the control group, with "00" (biased) patterns remaining dominant (71.4% in Study 1 and 85.3% in Study 2). However, in the training group, participants showed a clear decrease in "00" patterns (50.0% in Study 1 and 36.3% in Study 2), that was mostly compensated by a boost in "11" patterns (41.4% in Study 1 and 52.0% in Study 2), and seldom by a boost in "01" patterns (7.8% in Study 1 and 9.9% in Study 2). The higher proportion of "11" patterns after the intervention compared to the proportion of "01" patterns shows that the training improved intuitive reasoning. Accordingly, in the training group, most final correct responses were also initially correct. This finding highlights that the training helped participants intuit the correct solution strategy rather than correct an initial "erroneous" response through deliberation.

3.2.3. *Individual level directions of change classification*

To explore further how participants solved the problems, we performed an individual level accuracy analysis (Raoelison & De Neys, 2019) for each participant, on each conflict trial, from start to end of the experiment. This allowed us to observe in detail how the participants' response patterns evolved after the intervention.

By and large, Fig. 4 suggests that we can, as in Raoelison and De Neys (2019), roughly classify the participants in three main groups. First, participants who predominantly provide incorrect responses (i.e., "00"

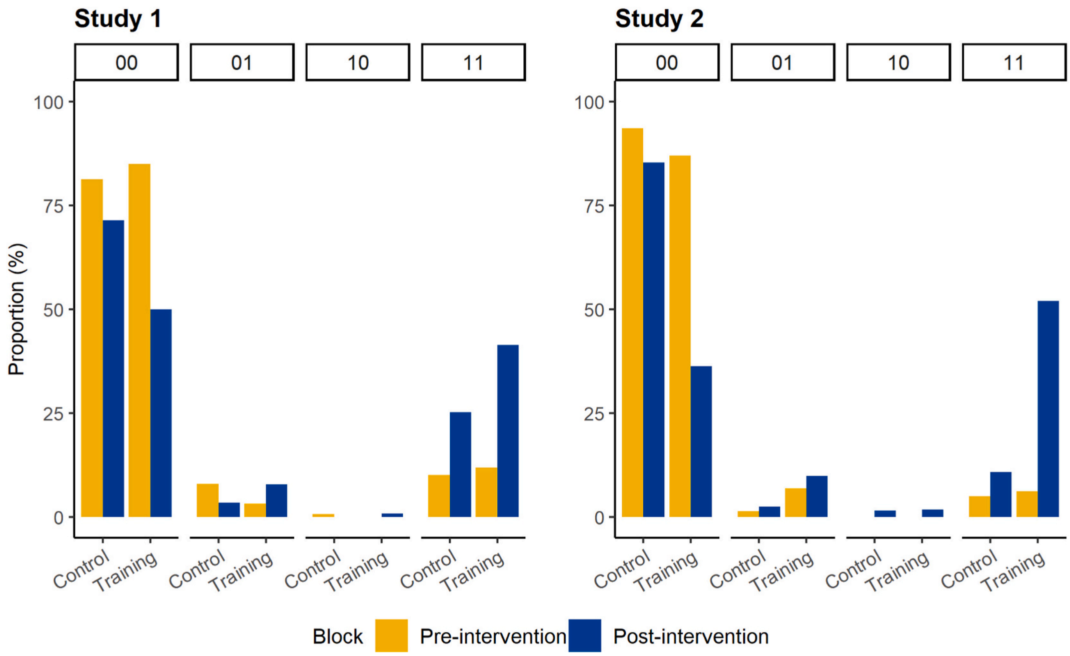

Fig. 3. Proportion of each direction of change (i.e., 00 trials, 01 trials, 10 trials and 11 trials) for the conflict problems according block and group in Study 1 and 2.

trials) before and after the intervention are labelled as "biased" respondents. These participants gave incorrect responses throughout the study and represent 73.3% of the participants in the control group and 43.6% of the participants in the training group, in Study 1, and 86.1% of the participants in the control group and 31.9% of the participants in the training group, in Study 2. Second, some reasoners already provide correct responses ("01" or "11" trials) in the pre-intervention block. These reasoners are labelled as "correct" respondents. These participants did not require any training intervention to respond correctly to bat-and-ball problems. In Study 1 and 2, respectively, 22.2% and 7% of the participants fell into this category in the control group, and 15.4% and 10.6% of the participants fell into this category in the training group. Third, some participants started by giving incorrect responses ("00" trials) and then, switched to correct responses ("01" or "11" trials) at some point after the intervention. This was rare in the absence of training in control group, 4.44% and 7% of participants in Study 1 and 2, respectively (these participants are referred to as "naturally improved" in Fig. 4). However, these "improved" participants represent 41.0% and 57.5% of the participants, in the training groups of studies 1 and 2, respectively.

An intergroup comparison showed that 94.3% (Study 1) and 92.5% (Study 2) of the pre-intervention biased respondents from the control groups remained biased after the (no-training) intervention. Only 5.7% and 7.5% of them were able to spontaneously provide at least one final correct response. In the training group, among the pre-intervention biased respondents, 48.5% and 64.3%, in studies 1 and 2 respectively, gave at least one final correct response after the training intervention. Again, this indicates that the training worked. Critically, among the improved reasoners' correct trials, 77.5% and 74.8% were of the "11" type (i.e., where both initial and final responses were correct). This again suggests that, when participants responded correctly after training, they typically intuited the correct solution and did no longer need to correct an erroneous "intuitive" response through deliberation.

Interestingly, the individual trial sequences in Fig. 4 also shows that although improved participants in Study 1 typically gave correct intuitive responses at the end of the post-intervention block, they frequently still erred at the start of the block. This indicates that reasoners might need a couple of trials to crystalize the insight they acquired during the training. In Study 2, we therefore tried to boost the training effect by giving participants an additional explanation during the intervention block (i.e., 3 vs 2 problems). In addition, we also added two extra problems to the post-intervention block to make sure that any possible later arriving correct responding was stable (e.g., whether correct responses in trial 4 were further followed by correct responses). As Fig. 4 shows, correct responding indeed occurred much sooner in the post-intervention sequence in Study 2—typically at the first or second trial.

3.2.4. Conflict detection

Previous studies have shown that, despite giving an incorrect response, reasoners sometimes sense their error or the presence of a response conflict (Bago & De Neys, 2017, 2019; De Neys, 2013; Frey, Bago, & De Neys, 2017; Hoover & Healy, 2019; Johnson et al., 2016; Mata, 2019; Pennycook, Fugelsang, & Koehler, 2015, but see also Mata, Ferreira, Voss, & Kollei, 2017). For instance, biased reasoners may doubt that their response is correct, as indicated by a decrease in response confidence when responding to conflict versus no-conflict problems. In this study, we explored whether the training intervention affected biased reasoners' conflict detection. That is although the training might not have managed to get biased people to reason accurately, it might have helped them to better detect that their answer is questionable. We used the typical conflict-detection index, by contrasting confidence[3] ratings for incorrectly solved conflict problems to confidence ratings for correctly solved no-conflict problems. We compared this index of conflict detection before and after the intervention, for both the training and control groups. A higher difference value implies a larger confidence decrease when solving conflict items, which is believed to reflect a more pronounced conflict experience (Bago & De Neys, 2019; Pennycook et al., 2015).

As Table 1 indicates, in both Study 1 and 2, there is indeed a trend towards a higher detection index after the intervention in the training group, especially for the initial responses. This effect is not observed in the control group. For completeness, we analysed the data using

[3] Since it has been shown that the initial response latency is not a reliable measure for conflict detection (Bago & De Neys, 2017), we will only present the conflict detection associated with the confidence rates (the conflict detection associated with response latency can be found in the Supplementary Material Section D).

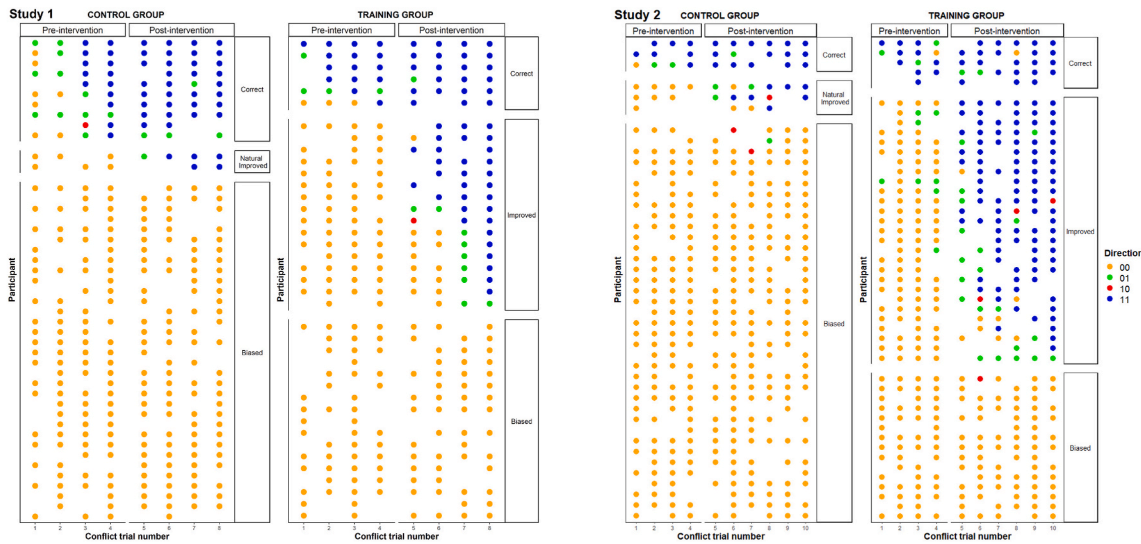

Fig. 4. Individual level direction of change (each row represents one participant) in Study 1 and 2. Due to discarding of missed deadline and load trials (see Trial Exclusion), not all participants contributed 8 analysable trials for Study 1 and 10 analysable trials for Study 2.

Table 1
Conflict detection results. Percentage of mean difference in confidence ratings (SE) between conflict and no-conflict problems.

Study	Group	Initial response		Final response	
		Pre-intervention	Post-intervention	Pre-intervention	Post-intervention
Study 1	Control	3.4% (2.3)	1.7% (3.0)	5.9% (2.7)	2.4% (3.1)
	Training	1.6% (5.0)	11.0% (7.3)	1.2% (4.8)	3.8% (8.2)
Study 2	Control	4.5% (1.6)	11.6% (3.3)	2.3% (2.2)	1.3% (4.1)
	Training	4.6% (3.9)	29.7% (6.9)	7.7% (4.4)	10.7% (4.8)

ANOVAs on the initial and final detection index with Block (pre- vs post-intervention) as a within-subjects factor and Group (training vs control) as a between-subjects factor. For the final responses, in Study 1 the ANOVA revealed no significant effect (All *Fs* < 0.059 and all *ps* $> .10$). In Study 2, the ANOVA revealed a trend for the Block x Group interaction, $F(1,58) = 3.0$, $p = .09$, $\eta^2 g = 0.02$, a main effect of Group, $F(1,58) = 6.1$, $p = .017$, $\eta^2 g = 0.06$, and no main effect of Block $F(1,58) = 1.1$, $p = .3$, $\eta^2 g = 0.01$. For initial responses, in Study 1, the ANOVA again failed to reveal any significant effect (all *Fs* $< 0.2.5$, and all *ps* $> .12$). In Study 2, the ANOVA revealed a significant Block x Group interaction, $F(1,66) = 15.8$, $p < .001$, $\eta^2 g = 0.09$, a main effect of Group, $F(1,66) = 10.6$, $p = .002$, $\eta^2 g = 0.09$, and a main effect of Block $F(1,66) = 15.2$, $p < .001$, $\eta^2 g = 0.09$.

In sum, the results suggest that, although some participants fail to provide the correct response after the training, they may nevertheless have benefited from it, in that they are better able to detect that their intuitive response may not be correct, at least in Study 2.

3.2.5. Predictive conflict detection

We also explored whether individual differences in one's ability to detect conflict (before the intervention) was predictive of the success of the intervention. That is, we examined whether the reasoners who started to respond correctly after the training intervention (i.e. improved respondents in our individual level classification) already showed better conflict detection before the training compared to those who did not improve after training (i.e. biased respondents). In order to do so, we compared conflict detection of improved vs biased respondents, before the training intervention, in the training group.

For final responses, in both studies 1 and 2, we observed a trend towards a better conflict detection in improved compared to biased respondents (Study 1: $t(31) = 1.4$, $p = .20$; Study 2: $t(30) = 1.9$, $p = .07$). The average conflict-detection rate was more pronounced for improved respondents (Study 1: M = 7.5%, SE = 7.4, Study 2: M = 10.1%, SE = 4.6) than for biased respondents (Study 1: M = −2.8%, SE = 7.4, Study 2: M = 2.1%, SE = 1.2). The same trend was observed for initial responses (Study 1: M improved = 7.5%, SE = 7.4; M biased = −1.2%, SE = 3.1; Study 2: M biased = 1.7%, SE = 4.8, M improved = 10.1%, SE = 4.6). The difference was not significant in Study 1, $t(31) = 0.5$, $p = .60$ while it showed a trend in Study 2: $t(40) = 1.8$, $p = .08$. Note that, for both initial and final responses, reasoners in the biased group did not show a nominal detection effect (i.e., the conflict detection index was negative), showing that these participants did not doubt their incorrect conflict responses.

3.2.6. Response latencies

Next, we explored participants' response latencies on the conflict problems. These were in line with previous two-response studies (e.g., Bago & De Neys, 2019). Overall, participants took slightly longer to respond in the final than in the initial response stage (Study 1: initial = 4.4 s, SE = 0.15, final = 7.5 s, SE = 0.68; Study 2: initial = 4.1 s, SE = 1.3, final = 7.6 s, SE = 0.60). For completeness, we ran a mixed-design ANOVA on the latencies using Block (pre- vs post-intervention) and Response-stage (initial vs. final) as within-subjects factors, and Group (training vs control) as a between-subjects factor. Fig. S1 in the Supplementary Material Section E shows the results. The analysis indicated that there was a significant effect of the Response Stage (Study 1: $F(1,81) = 23.46$, $p < .001$, $\eta^2 g = 0.08$; Study 2: $F(1,88) = 48.32$, $p < .001$, $\eta^2 g = 0.11$) and Block factor (Study 1: $F(1,81) = 7.01$, $p = .01$, $\eta^2 g = 0.01$; Study 2: $F(1,88) = 5.25$, $p = .02$, $\eta^2 g = 0.01$), indicating that participants responded overall faster in the initial than final response stage and faster in the post vs. pre-intervention stage. In Study 2, there was also a Group ($F(1,88) = 5.09$, $p = .03$, $\eta^2 g = 0.02$) and Group x Response Stage ($F(1,88) = 4.33$, $p = .04$, $\eta^2 g = 0.01$) interaction indicating that the longer final vs initial latencies were most pronounced in the Training group. However, this effect was already present in the pre-intervention and was not observed in Study 1. None of the other factors or interactions reached significance (all *Fs* < 1.53 and *ps* > 0.22). Hence, there was no clear evidence suggesting that the training intervention affected response times per se.

3.2.7. Transfer problem accuracy

We explored whether the training intervention led to an enhancement of performance on two types of untrained problems (CRT-like and neutral problems).

For CRT-like problems, as shown in Fig. 5, there was no effect, except for a general pre- to post-intervention increase in initial-response accuracy in both studies 1 and 2. The ANOVAs revealed that these improvement were similar across participants, whether they were trained or not (Study 1: main effect of Block, $F(1,74) = 13.9$, $p < .001$, $\eta^2g = 0.07$; no main effect of Group $F(1,74) = 1.2$, $p = .28$, $\eta^2g = 0.01$ and no significant Block x Group interaction, $F(1,74) = 0.3$, $p = .58$, $\eta^2g = 0.002$; Study 2: main effect of Block $F(1,77) = 11.9$, $p = .001$, $\eta^2g = 0.03$; no main effect of Group $F(1,77) = 0.1$, $p = .8$, $\eta^2g = 0.001$; nor a significant Block x Group interaction, $F(1,77) = 0.8$, $p = .37$, $\eta^2g = 0.002$). Likewise, final–response accuracy did not vary as a function of any of the independent variables in Study 1: Block, $F(1,79) = 1.2$, $p = .27$, $\eta^2g = 0.003$; Group $F(1,79) = 0.6$, $p = .46$, $\eta^2g = 0.006$ and their interaction $F(1,79) = 0.1$, $p = .8$, $\eta^2g = 0.0002$. In Study 2, only the main effect of Block was significant (main effect of Block $F(1,80) = 4.8$, $p = .03$, $\eta^2g = 0.01$; no main effect of Group $F(1,80)$ =0 0.6, $p = .44$, $\eta^2g = 0.01$; nor a significant Block x Group interaction, $F(1,80) = 0.1$, $p = .75$, $\eta^2g = 0.00$). In sum, the training intervention did not yield any transfer to CRT-like problems.

We also wanted to test whether the training could lead to an enhancement of performance on simple neutral arithmetic word problems. Fig. 5 shows the results. However, as with the CRT-like problems, Fig. 5 indicates that except for a general pre- to post-intervention increase in accuracy, there was no clear sign of a training effect on the neutral arithmetic problems. Analysis-wise, in Study 1,[4] for both response stages (i.e., initial and final), we found that Block significantly improved the model fit (Initial response: $\chi2$ (1) = 7.54, $p = .01$; Final response: $\chi2$ (1) = 6.84, $p = .01$) but not Group (Initial response: $\chi2$ (1) = 0.34, $p = .56$; Final response: $\chi2$ (1) = 0.03, $p = .86$), nor their interaction (Initial response: $\chi2$ (1) = 0.50, $p = .48$; Final response: $\chi2$ (1) = 0.01, $p = .93$). In Study 2, for both response stages (i.e., initial and final), the ANOVA showed no interaction of Group x Block (Final; $F(1, 80) = 0.5$, $p = .48$, $\eta^2g = 0.02$ and Initial; $F(1,73) = 1.9$, $p = .17$, $\eta^2g = 0.01$), nor a main effect of Group (Final; $F(1, 80) = 1.3$, $p = .25$, $\eta^2g = 0.01$ and Initial; $F(1,73) = 0.1$, $p = .8$, $\eta^2g = 0.001$). There was a main effect of Block $F(1,73) = 11.6$, $p = .001$ $\eta^2g = 0.03$, for initial responses but not for final responses $F(1,80) = 1.0$, $p = .33$, $\eta^2g = 0.003$.

In sum, both on the CRT-like and neutral transfer problems, participants tended to improve somewhat when they solved the problems a second time in the post-intervention phase, but this improvement was not specifically boosted by the training. Hence, the results suggest that the training effect is highly specific to the bat-and-ball problem and does not lead to an overall increase in performance on other, untrained reasoning tasks.

3.2.8. Bat-and-two-balls problem accuracy

Studies 1 and 2 showed that training people on the bat-and-ball problem helps them to intuit the correct answer on this specific problem (but not on others). A possible critique to our study is that our explanations did not help reasoners to grasp the underlying bat-and-ball problem logic but simply let participants develop an alternative "heuristic" shortcut. For example, in theory, one possibility is that participants simply rote memorize the correct response ("It's 5 cents"). Clearly, given that all our training and test blocks used content-modified items with unique quantities, such a confound is readily ruled out. A more realistic concern is that participants develop some sort of "halving heuristic" ("It's always half of what you think it is!") in which they blindly half the cued original heuristic response. This version is ruled out by our control problems. Here the cued heuristic response is also correct, and performance was near ceiling. If participants engaged in blind "halving", they should have massively erred here. However, a more advanced 'selective' version of this heuristic would note, for example, that the control problems do not contain the word "more". Hence, participants would only use halving if they see the "more" cue (e.g., "If 'more', than take half of what you think"). As with the control problems, findings on the neutral problems argue against this confound. Neutral problems also contain the "more" statement, and although we did not observe a transfer effect, initial accuracy after training hovered around 75%.

Nevertheless, one may note that the neutral problems still have a different underlying structure (e.g., they do not contain "more *than* X", do not cue a heuristic response, etc.) that might be used as an advanced selective halving cue. In Study 2 we therefore created new "bat-and-two-balls" problems ("A bat and two balls cost $2.60 in total. The bat costs $2 more than two balls. How much does one ball cost?"). They require an additional division, but the basic underlying structure and substitution logic is completely similar to the original bat-and-ball logic. If reasoners simply use halving, they will err ("30 cents"), but if reasoners understand the logic after training, correct bat-and-two-balls answers ("15 cents") should also increase.

Fig. 6 provides an overview of the average performance of the training and control groups. First, we focus on final-response accuracies. Most reasoners, from both the training and control groups, failed to solve the bat-and-two-balls problems before the intervention (respectively, M = 14.8%, SE = 5.0, and M = 6.4%, SE = 3.8). Both groups improved in average performance after the intervention, but the improvement was larger for the training group (overall accuracy increase of 27.3%, SE = 13.3) than for the control group (overall accuracy increase of 7.7%, SE = 5); the Block x Group interaction was significant $F(1,81) = 5.6$, $p = .02$, $\eta^2g = 0.02$. The training intervention led participants to produce more final correct responses for the bat-and-two-balls problems.

We also tested whether the training effect occurred for initial "intuitive" responses. Before the intervention, in both groups, most of the participants failed to solve the bat-and-two-balls problems (Training group: M = 5.7%, SE = 2.9; Control group: M = 6.4%, SE = 3.8). After the intervention, while overall performance increased in both groups, it increased more in the training group (overall increase of 22.7%, SE = 5.6) than in the control group (overall increase of 6.41%, SE = 4.2); the Block x Group interaction was significant, $F(1,81) = 5.36$, $p = .02$, $\eta^2g = 0.02$.

To further control for a possible "halving confound" we also explored how the prevalence of the "halving" response on the bat-and-two-balls problems changed after training. We therefore separated participants in the training condition in three groups according to their accuracy patterns ('correct', 'improved', and 'biased'; see above). Results indicate that participants who benefited from the training (i.e., the improved group) gave more correct and *fewer* halving responses after training. Interestingly, if anything, it was only the subjects whose performance did not increase (i.e., biased respondents) who tended to start using the halving strategy more after training (see Supplementary Material Section F for full overview). This establishes that our observed increased initial response accuracy does not result from a halving confound.

4. Study 3

Studies 1 and 2 showed that a short training on the bat-and-ball-problem can help people to intuit the correct response. With Study 3, we aimed to test whether the training effect sustained over time. In order to do so, two months after completion of Study 2, trained participants of that study were invited to take part in a re-test, (i.e., Study 3). Study 3 used the same procedure as Study 2 (except that all problems had a different surface content). After the pre-intervention block, participants again went through our training intervention and completed a post-

[4] In Study 1, due to a coding error, one of the two neutral-problem responses was not recorded in the post-intervention block. Because our data were composed of binary responses, we applied a mixed-effect logistic regression in which participants were entered as random effect intercept for those data.

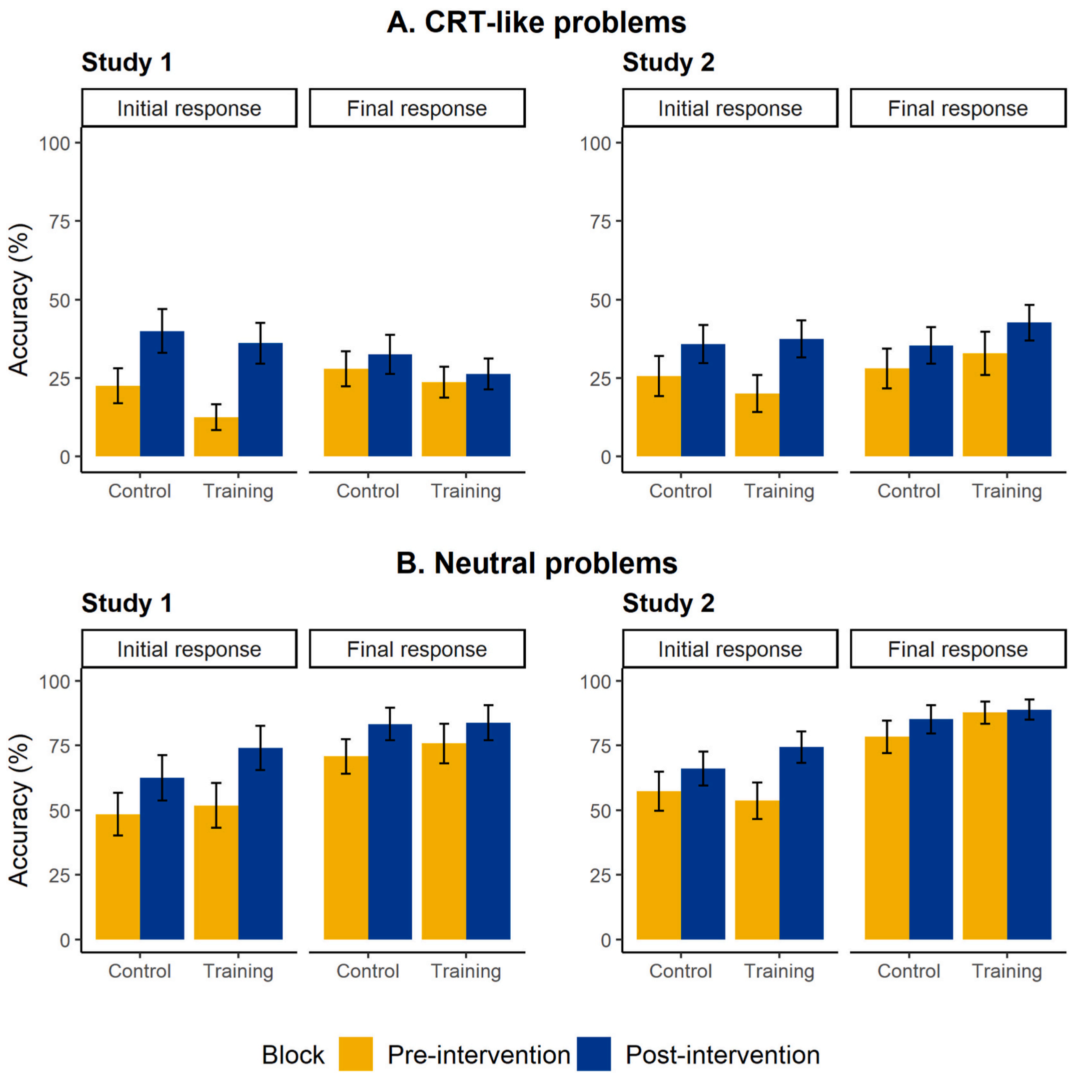

Fig. 5. Average initial and final accuracy on CRT-like (panel A) and neutral problems (panel B) in Study 1 and 2. Error bars represent standard error.

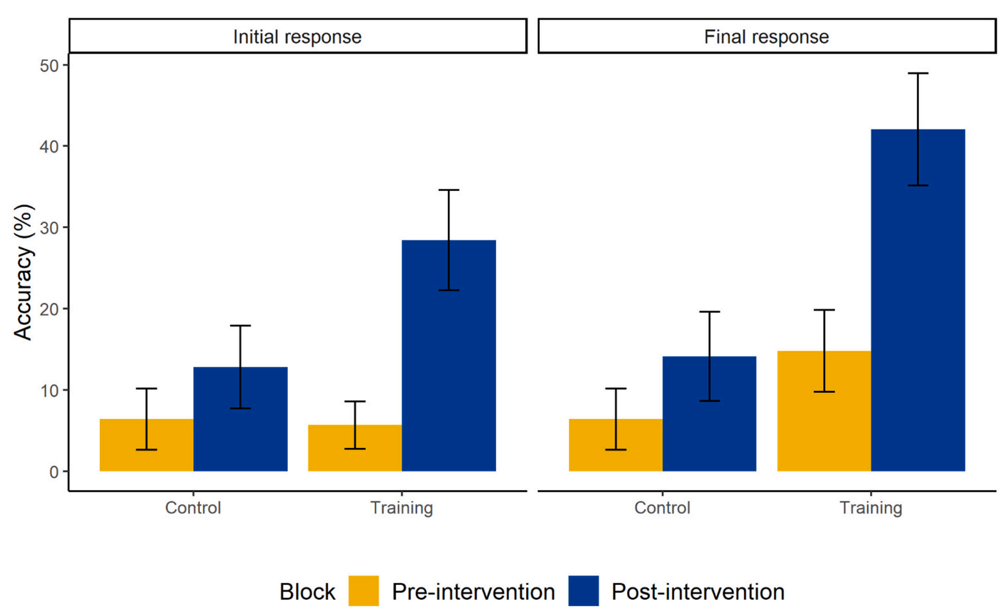

Fig. 6. Average initial and final accuracy on bat-and-two-balls problems. Error bars are standard errors.

intervention block. This also allowed us to explore whether giving participants an additional training session could further boost performance.

4.1. Methods

4.1.1. Preregistration

The study design and hypothesis were preregistered on the Open Science Framework (http://osf.io/qx7fc). No specific analyses were preregistered.

4.1.2. Participants

Thirty-four participants took part in Study 3 (out of the 47 participants in the Study 2 training group; 26 females, M = 33.36 years, SD = 10.83). One of them only completed the pre-intervention block. The sample was composed of nine people who were classified as biased respondents in Study 2, three were correct respondents and 22 were improved respondents. We compensated participants for their time at the rate of £5 per hour.

4.1.3. Materials and procedure

The material and the procedure were the same as in Study 2. All the problems featured modified contents (see Supplementary Material Section A).

4.1.4. Trial exclusion

Participants failed to provide their first answer before the deadline on 28 trials (2.7% of all trials) and failed to pick the correct matrix on the load task on 123 trials (12.4% of the remaining trials). We discarded these trials and analysed the remaining 869 trials (85.2% of all trials). On average, each participant responded to 25.5 (SD = 4.1, max number trials = 30) trials.

4.2. Results and discussion

4.2.1. The sustained training effect

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

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

First, we focus on final-response accuracies. Fig. 7 shows that, while participants gave fewer correct responses two months after training (in the pre-intervention block of Study 3; M = 51.5%, SE = 8.4) than just after training (in the post-intervention block of Study 2; M = 66.4%, SE = 7.4), $t(33) = 2.3$, $p = .03$, they nevertheless gave more correct responses two months after training than before their first training (in the pre-intervention block of Study 2; M = 12.26%, SE = 5.0), $t(33) = 5.19$, $p < .001$.

The same trend was observed with initial responses. Despite a decrease in performance observed two months after training (Study 3 pre-intervention: M = 40.0%, SE = 7.5), compared to just after training (Study 2 post-intervention: M = 57.8%, SE = 7.1), $t(33) = 2.9$, $p = .008$, performance clearly remained better than before the first training (M = 6.4%, SE = 3.7), $t(33) = 4.8$, $p < .001$.

In Study 3, we managed to reach 72% (34/47) of the Study 2 participants. To check for a possible attrition confound (e.g., subjects who did better in Study 2 were more likely to sign-up for Study 3), we compared the Study 2 pre-intervention conflict problem accuracy of the

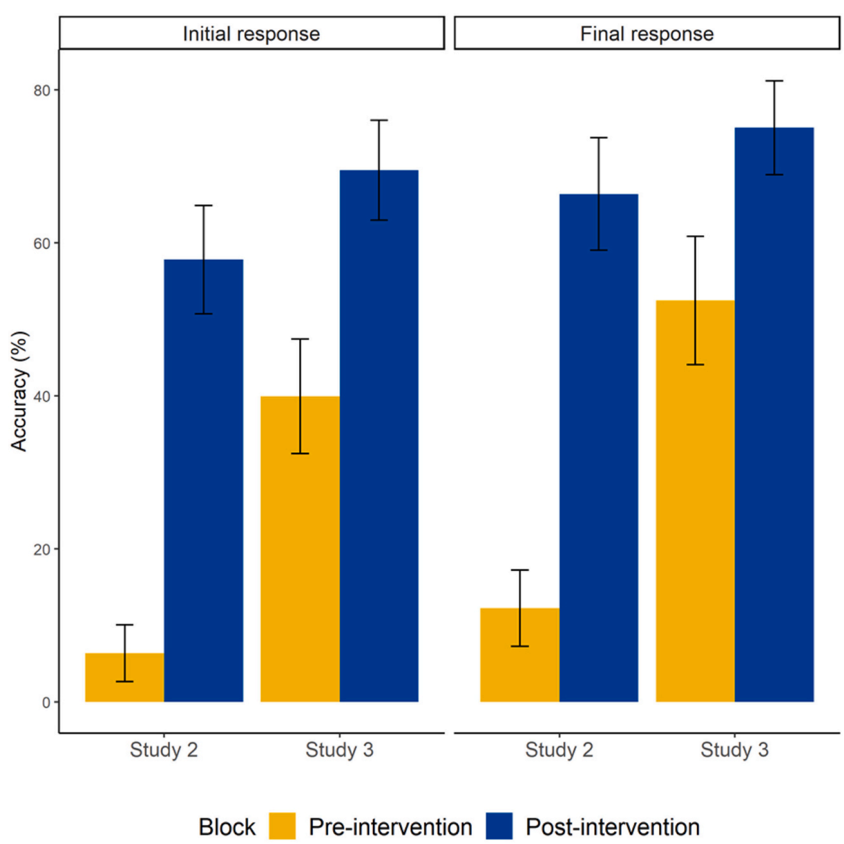

Fig. 7. Average initial and final accuracy on conflict problems in Study 2 (pre- and post-intervention) and Study 3 (pre- and post-intervention). Error bars are standard errors.

subgroup of Study 3 participants (Initial response: M = 6.4%, SE = 3.7; Final response: M = 12.3%, SE = 5.0) to the overall Study 2 pre-intervention conflict problem accuracy (Initial response: M = 8.3%, SE = 3.7; Final response: M = 15.3%, SE = 4.7). Given that our Study 3 participants did not score better than the Study 2 average, it is unlikely that the Study 3 results are artificially boosted because of an attrition confound.

In conclusion, the training intervention effect was robust and sustained over time, for at least two months, for both initial 'intuitive' responses and final 'deliberate' responses. This result was also backed up by a direction of change analysis (see Supplementary Material Section G).

For completeness, no-conflict problem accuracies were also analysed. Despite a slight decrease in performance two months after the intervention, for both final and initial responses, performance remained near ceiling (see Supplementary Material Section C).

4.2.1.2. Individual level directions of change classification. To get a more detailed picture, Fig. 8 shows the proportion of each direction of change in Studies 2 and 3 separately for those reasoners who were classified as Biased, Correct and Improved respondents based on the Study 2 classification. A visual inspection of the data shows that correct respondents (i.e., reasoners who answered correctly before receiving any training, n = 3) kept giving a majority of "11" response patterns two months after training, while biased respondents (i.e., reasoners who were still biased after the Study 2 training, n = 9) remained biased two months later, mainly giving "00" response patterns. In comparison, improved respondents (i.e., reasoners who benefitted from training, n = 22) gave more "00" response patterns two months after the training intervention (28.4%) than just after it (8.2%), but far less than before training (93.6%). In addition, improved respondents produced more "01" and "11" response patterns (respectively 22.4% and 47.0%) than just before training (respectively, 6.4% and 0%). Critically, even two months after the intervention, they were still more likely to produce "11" response patterns (47.0%) than "01" response patterns (22.4%), suggesting that the training provided in Study 2 led most participants to intuit the correct solution strategy (rather than correcting an "erroneous" intuition) over a period of at least two months. In sum, the results suggest that the training effect persisted over time for those who improved in performance after the training intervention of Study 2.

4.2.1.3. Additional data. For completeness, consistent with Study 2, we also presented additional transfer and neutral problems, collected confidence ratings and justifications. We had no a priori hypotheses about these data but the interested reader can find an overview in the Supplementary Material (Section B for justifications, Section H for CRT-like problems, Section I for neutral problems and Section J for the conflict detection). Study 3 also included the bat-and-two-balls problems. The full analysis can also be found in the Supplementary Material Section K. We simply note here that as with the standard bat-and-ball problems the initial bat-and-two-balls accuracy decreased in Study 3 but was still higher than before the training. This indicates that the sustained bat-and-ball performance was not driven by an increased application of the halving heuristic per se.

4.2.2. Second training effect

We also tested whether a second training (i.e., in Study 3) could further improve the performance. We compared performance across the pre- and post-intervention blocks of Study 3, and across the post-intervention blocks of Study 2 and of Study 3.

4.2.2.1. Bat-and-ball response accuracy. First, we focus on final-response accuracies. Fig. 7 shows that participants gave more correct responses after the training intervention of Study 3 (M = 75.01%, SE = 6.1) than just before it (M = 51.0%, SE = 8.5), $t(32) = 3.8$, $p < .001$. However, the difference between Study 3 post-intervention performance (M = 75.0%, SE = 6.1) and Study 2 post-intervention performance (M = 65.4%, SE = 7.5) did not reach significance, $t(32) = 1.6$, $p = .12$, suggesting that the increase in performance after the second training was only marginal.

With respect to initial-response accuracy, participants' performance was again higher after the training intervention of Study 3 (M = 69.5%, SE = 6.5) than just before it (M = 38.9%, SE = 7.63), $t(32) = 5.2$, $p <$ .001, and than after the training intervention of Study 2 (M = 56.5%, SE = 7.2), $t(32) = 2.4$, $p = .02$. Hence, the slight performance decrease two months after Study 2 was remediated with an additional training, and this training even helped to go beyond the initial Study 2 training performance. The accuracy results were also backed up by a direction of change analysis (see Supplementary Material Section G).

No-conflict problem accuracies can be found in Supplementary Material Section C. Performance was near ceiling for both final and initial responses.

4.2.2.2. Individual level directions of change classification. We also performed a direction of change analysis according to the type of respondent classification in Study 2. Mirroring the overall accuracy effects, in

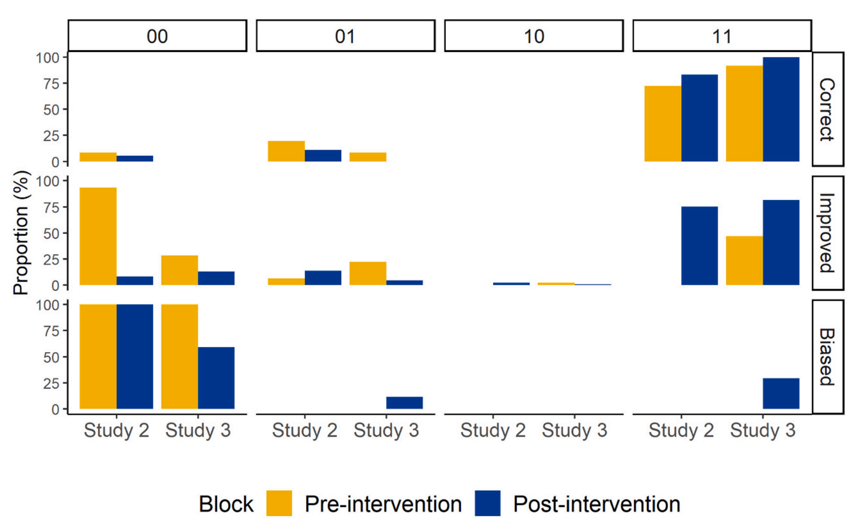

Fig. 8. Proportion of each direction of change (i.e., 00 trials, 01 trials, 10 trials and 11 trials) for the conflict problems according to block and type of respondents in Study 2 and 3.

both the “correct-respondent” and “improved-respondent” groups, the proportion of “11” response patterns reached its highest level after the second training, compared to just before it, and compared to after the first training (see Fig. 8). More importantly, among the biased respondents of Study 2, who had not yet shown competency in solving bat-and-ball-like problems, we started to observe correct answers (11.5% “01” and 29.3% “11”) after the second training (in Study 3). Critically, as the proportion of “11” trials suggests, such correct answers were often already generated intuitively. This tentatively suggests that repetitive training might allow even more individuals to intuit the correct solution strategy.

5. General discussion

The present study explored whether we can de-bias reasoners and boost correct intuitive responses with a short training intervention. We ran three studies using a two-response protocol in which participants were asked to provide two consecutive responses—one initial “intuitive” and one final “deliberate”—to adaptations of the bat-and-ball problem. Consistent with other studies, the findings indicated that training led a majority of biased participants to improve their performance. Critically, we found that training enabled most reasoners to give a correct answer as early as the intuitive stage. After the training, participants no longer needed to deliberate to correct their intuition and this sound intuiting effect was observed up to two months after the first training.

The results indicate that once people are told how to solve the problem, they can quickly automatize the application of the underlying mathematical operations and generate correct responses without any further deliberation. At a more theoretical level this helps to provide some insight into the nature of the bat-and-ball errors. The training results make it crisp clear that (at least for the modal reasoner, see further) the bias results from a performance rather than a competence error (e.g., Hoover & Healy, 2017, 2019; Mata, 2020). The problem is not that people do not know the necessary underlying logico-mathematical operations but rather that they are not using their knowledge. Obviously, it would be ludicrous to argue that a five-minute training suffices to learn the underlying algebraic equation logic ex nihilo. That is, the fact that the short explanation worked and allowed people to intuit correctly suggests that all the critical building blocks were already there. Indeed, all educated adults have been taught how to solve similar equations and practiced the operations at length in their high school math courses (Hoover & Healy, 2017). Hence, once the problem structure is clarified and the relevance of the building blocks becomes clear, correct responding can become a “no brainer”. Against this backdrop, the results should be less surprising than they may be at first sight perceived by some. People can intuitively perform the necessary operations precisely because they have long acquired and (to some extent) automatized them. The implicit knowledge is already there, people simply need to be reminded how to put it to use.

The finding that de-biasing training can actually help people intuit correctly, has also important applied implications. Traditionally, it is often assumed that de-biasing interventions work by boosting deliberation and get people to better correct erroneous intuitions (Lilienfeld, Ammirati, & Landfield, 2009; Milkman, Chugh, & Bazerman, 2009). As we noted in the introduction, although it can be laudable to help people to deliberate more, in many daily life situations they will simply not have the time (or resources) to successfully deliberate. Hence, if de-biasing interventions only help people to deliberate more, their impact may be limited. Ultimately, we do not only want people to correct erroneous intuitions but to avoid biased intuitions altogether (Evans, 2019; Milkman et al., 2009; Reyna, Weldon, & McCormick, 2015; Stanovich, 2018). What the present study indicates is that existing de-biasing interventions, in which the problem logic is briefly explained, might be more powerful in this respect than hitherto assumed.

Given the potential theoretical and applied impact of the findings, it is important to avoid possible misconceptions and keep limitations in mind. One possible critique to our study is that our training explanations did not help reasoners to grasp the underlying bat-and-ball problem logic but simply cued participants to use an alternative “heuristic” shortcut (e.g., “it’s half of what you think it is”). The high accuracies on our control no-conflict and neutral problems together with our findings on the bat-and-two-balls problems argue against such a simple confound. The latter problems were designed to share the same underlying equation logic but simply required an additional division. If participants understand the underlying bat-and-ball structure, they should also manage to solve the bat-and-two-balls problems. Results showed that successful training also boosted correct intuiting on the bat-and-two-balls problems whereas erroneous “halving” responses did not increase. Taken together these results present good evidence against a possible “halving” heuristic confound.

To avoid confusion, it should be stressed that our bat-and-two-balls problems were explicitly designed to share the underlying bat-and-ball structure. Results on our proper transfer tasks clearly showed that the training effect did not generalize to other non-trained reasoning tasks. Neither people’s performance on basic algebraic word problems, nor CRT-like lure problems was specifically enhanced after training. This indicates that reasoners did not intuit *(or deliberate)* better in general. They got better at solving the very specific problem they were explained. This fits with the finding that existing de-biasing or cognitive training programs are often task or domain specific (Lilienfeld et al., 2009; Sala & Gobet, 2019; but see also Morewedge et al., 2015; Trouche et al., 2014).

Note that, at the practical side, the task-specificity of trained sound intuiting does not necessarily present a drawback. The actual training intervention took less than five minutes and did not require any intervention from a human teacher. Hence, the costs (both in terms of time and resources) are minimal. Instead of having reasoners go through a (lengthy) generic training which is hoped to transfer, one could envisage giving them a battery of short task-specific interventions that are each designed to focus on one specific problem. Although speculative, the lack of training transfer would be less problematic than it could be perceived in this respect.

As a side note, we tentatively speculate that the task specificity might be an intrinsic feature of training interventions aimed at the “System 1” level. The intuitive System 1 has long been characterized as more domain specific than the deliberate System 2 (Reber, 1992). Intuitions can be conceived as a highly specialized set of procedures that have been practiced to automaticity and are autonomously executed when their triggering stimulus is encountered (e.g., Stanovich, 2009). Under this view, our training might help to boost the mapping between a specific problem structure X and operation Y. However, the mapping between an alternative problem structure W and operation Z will obviously not be affected. Hence, the point we try to highlight is that simply because of the nature of intuitive or automatized reasoning procedures, transfer might be necessarily limited.

It should be clear that our results do not argue against a role of deliberation in de-biasing per se. Our key finding is that once the bat-and-ball is briefly explained to reasoners, they can readily automatize the required operations and intuit correctly. But the fact that people no longer need to deliberately correct once they grasp the solution strategy does not mean that deliberation plays no role in achieving this understanding per se. For example, during our intervention block in which the problem was explained to reasoners, they were not under time or dual task pressure and could take all the time they wanted to reflect on the explanations. Indeed, if one wants to explain a problem, it would be nonsensical to not let people reflect on it. This role of deliberation in helping people understand the problem structure is also illustrated by the fact that in our post-intervention block, many participants generate at least one deliberate corrective trial (i.e., “01”) before they start giving intuitive correct responses. Hence, the first time they show “insight” typically happens during deliberation. In sum, our point is not that people do not need to deliberate to understand how to solve the bat-and-ball problem. The point is that once people understand this, they also readily automatize the proper operations and no longer need to deliberate to correct their intuition.

When we state that training helps biased reasoners to intuit correctly it is important to keep in mind that we are talking about the modal (or average) reasoner. Our results show that the majority of biased reasoners learned to intuit correctly after training. However, there were also individual exceptions. Some individuals remained biased after training. Interestingly, we found that the training effect tended to be predicted by participants' spontaneous conflict or error detection. Biased reasoners who became more accurate after training showed more conflict detection (i.e., doubted their incorrect answer more) *before* the training than those who did not improve. Hence, it seems that they had a more advanced knowledge state than those who failed to benefit from the training. Although they did not manage to intuit the correct answer spontaneously, they at least seemed to realize their heuristic answer was questionable.

Our conflict detection analysis further indicated that even for reasoner who remained biased, the training was not completely unsuccessful. After training, incorrect responders tended to doubt their erroneous answers more than before the training. Hence, although the training did not help them to answer correctly yet, it at least seemed to help them realize that their erroneous response was not fully warranted. Interestingly, Study 3 indicated that with repeated training we also started to observe some correct intuiting among these reasoners. Although speculative, this suggests that even these reasoners might have the necessary competence or "building blocks" to solve the problem but their knowledge is less instantiated or activated (e.g., Stanovich, 2018). Hence, with more extensive training they might be brought up to the level of spontaneous sound reasoners.

We believe that the present study can serve as a proof-of-principle that underscores the potential of training sound intuiting. We focused on the-bat-and-ball problem because it is one of the most notorious examples of biased reasoning, which the majority of educated adults fail to solve spontaneously. Indeed, it has sometimes been questioned whether people can be properly de-biased on this problem in the first place (Bourgeois-Gironde and Van Der Henst, 2009). The fact that a simple intervention manages to get the majority of biased reasoners to intuit correctly is clearly noteworthy in this respect. Nevertheless, our study is but the first in which the issue is empirically explored. It will be important to validate and fine-tune the present findings. For example, although the trainability of a problem as notorious as the bat-and-ball problem is promising, it will be important to test the generalizability towards other reasoning tasks. In the first place, one can envisage generalization towards the classic logico-mathematical bias tasks that have long been studied in the reasoning and decision-making literature (e.g., Kahneman, 2011). However, biased reasoning is hurting performance in a very wide range of more applied contexts. One might think here, for example, of classroom settings (e.g., Beaulac & Kenyon, 2018; Brault Foisy, Ahr, Masson, Borst, & Houdé, 2015; Brault Foisy, Matejko, Ansari, & Masson, 2020), sharing of fake news on social media (Bago, Rand, & Pennycook, 2020; Pennycook & Rand, 2019), fixation effects in engineering design (e.g., Agogué, Le Masson, Dalmasso, Houdé, & Cassotti, 2015), machine algorithm aversion (e.g., Baer, 2019; Bonnefon, Shariff, & Rahwan, 2016), gender discrimination in hiring (e.g., Isaac, Lee, & Carnes, 2009), or racial biases in policing decisions (e.g., Payne, 2006). Ideally, future studies should also test the trainability of sound intuiting in these settings.

Relatedly, it is plausible that the efficacy of the training can be further optimized. Study 3 indicated that additional training helped at least some biased reasoners to improve. One can, for example, envisage how repeating the training on a number of consecutive days might further boost its efficacy. Obviously, the optimal approach remains to be explored here. In other words, we see the study as a critical proof-of-principle and starting point. Various scholars have pointed out the importance and theoretical possibility of training sound intuiting (or "System 1" training, e.g., Evans, 2019; Milkman et al., 2009; Stanovich, 2018; Reyna et al., 2015). The present study indicates that this is not a naïve utopian promissory note. We hope that theorists and practitioners will take note, and this will lead the field towards a deeper empirical exploration of sound intuiting in the coming years. De-biasing our "System 1" might be more straightforward than many have traditionally assumed.

Data availability

Raw data can be downloaded from our OSF page (https://osf.io/qx7fc/).

Declaration of Competing Interest

None.

Acknowledgments

This research was supported by a research grant (DIAGNOR, ANR-16-CE28-0010-01) from the Agence Nationale de la Recherche, France.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104645.

References

Agogué, M., Le Masson, P., Dalmasso, C., Houdé, O., & Cassotti, M. (2015). Resisting classical solutions: The creative mind of industrial designers and engineers. *Psychology of Aesthetics, Creativity, and the Arts, 9*, 313–318. https://doi.org/10.1037/a0039414.

Baer, T. (2019). *Understand, manage, and prevent algorithmic Bias: A guide for business users and data scientists*. Apress.

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

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

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General, 149*(8), 1608–1613. https://doi.org/10.1037/xge0000729.

Bago, B., Raoelison, M., & De Neys, W. (2019). Second-guess: Testing the specificity of error detection in the bat-and-ball problem. *Acta Psychologica, 193*, 214–228. https://doi.org/10.1016/j.actpsy.2019.01.008.

Beaulac, G., & Kenyon, T. (2018). The scope of debiasing in the classroom. *Topoi, 37*(1), 93–102. https://doi.org/10.1007/s11245-016-9398-8.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654.

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

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

Brault Foisy, L.-M., Matejko, A. A., Ansari, D., & Masson, S. (2020). Teachers as orchestrators of neuronal plasticity: Effects of teaching practices on the brain. *Mind, Brain, and Education, 14*(4), 415–428. https://doi.org/10.1111/mbe.12257.

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. (2013). Heuristics, biases, and the development of conflict detection during reasoning. In H. Markovits (Ed.), *Development of reasoning*. Hove, UK: Psychology Press.

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.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist, 16*. https://doi.org/10.1037/0003-066X.49.8.709.

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

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

Evans, J. S. B. 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.

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

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

Hoover, J. D., & Healy, A. F. (2019). The bat-and-ball problem: Stronger evidence in support of a conscious error process. *Decision, 6*, 369–380. https://doi.org/10.1037/dec0000107.

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

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

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. (2005). A model of heuristic judgment. In K. J. Holyoak, & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge University Press.

Lawrence, M. A. (2016). ez: Easy analysis and visualization of factorial experiments [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=ez.

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

Mata, A. (2019). Conflict detection and social perception: Bringing meta-reasoning and social cognition together. *Thinking & Reasoning*, 1–10. https://doi.org/10.1080/13546783.2019.1611664.

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

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: An attentional account of reasoning errors. *Psychonomic Bulletin & Review, 24*(6), 1980–1986. https://doi.org/10.3758/s13423-017-1234-7.

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

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

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

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights From the Behavioral and Brain Sciences, 2*(1), 129–140. https://doi.org/10.1177/2372732215600886.

Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1154–1170. https://doi.org/10.1037/xlm0000372.

Payne, B. K. (2006). Weapon bias: Split-second decisions and unintended stereotyping. *Current Directions in Psychological Science, 15*(6), 287–291. https://doi.org/10.1111/j.1467-8721.2006.00454.x.

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.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011.

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., Keime, M., & De Neys, W. (2021). *Think slow, then fast: Does repeated deliberation boost correct intuitive responding?* Memory & Cognition.

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

R Core Team. (2017). *R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria*. Retrieved from https://www.R-project.org/.

Reber, A. S. (1992). The cognitive unconscious: An evolutionary perspective. *Consciousness and Cognition, 1*(2), 93–133. https://doi.org/10.1016/1053-8100(92)90051-B.

Reyna, V. F., Weldon, R. B., & McCormick, M. (2015). Educating intuition: Reducing risky decisions using fuzzy-trace theory. *Current Directions in Psychological Science, 24*(5), 392–398. https://doi.org/10.1177/0963721415588081.

Sala, G., & Gobet, F. (2019). Cognitive training does not enhance general cognition. *Trends in Cognitive Sciences, 23*(1), 9–20. https://doi.org/10.1016/j.tics.2018.10.004.

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. N., Pennycook, G., & Rand, D. G. (2018). Performance on the cognitive reflection test is stable across time. *Judgment and Decision making, 9*. https://doi.org/10.1177/0146167218783192.

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. Evans & K. Frankish (Éds.), Two minds: Dual processes and beyond. Oxford University Press. 55–88 doi: https://doi.org/10.1093/acprof:oso/9780199230167.001.0001.

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

Stanovich, K. E. (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.

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

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). https://doi.org/10.1371/journal.pone.0186404.

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.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making, 11*(1), 15.

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.

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. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. http://ggplot2.org.

Wickham, H., Francois, R., Henry, L., & Muller, K. (2020). Dplyr: A grammar of data manipulation [R package version 0.8.5]. https://CRAN.R-project.org/package=dplyr.

Wickham, H., & Henry, L. (2020). Tidyr: Tidy messy data [R package version 1.0.2]. https://CRAN.Rproject.org/package=tidyr.