"You're Wrong!" : The Impact of Accuracy Feedback on the Bat-and-Ball Problem

Eva M. Janssen[a], Matthieu Raoelison[b], Wim De Neys[b]
[a] Utrecht University, The Netherlands
[b] Université de Paris, LaPsyDE (UMR CNRS 8240), France

Abstract

The popular bat-and-ball problem is a relatively simple math riddle on which people are easily biased by intuitive or heuristic thinking. In two studies we tested the impact of a simple but somewhat neglected manipulation – the impact of minimal accuracy feedback – on bat-and-ball performance. Participants solved a total of 15 standard and 15 control versions of the bat-and-ball problem in three consecutive blocks. Half of the participants received accuracy feedback in the intermediate block. Results of both studies indicated that the feedback had, on average, no significant effect on bat-and-ball accuracy over and above mere repeated presentation. We did observe a consistent improvement for a small number of individual participants. Explorative analyses indicated that this improved group showed a more pronounced conflict detection effect (i.e., latency increase) at the pretest and took more deliberation time after receiving the negative feedback compared to the unimproved group.

Keywords: reasoning; heuristics and biases; bat-and-ball problem; feedback; learning; conflict detection.

Introduction

Although humans have unique cognitive abilities to reason, human reasoning can also be biased. For instance, investors can make bad investment decisions based on the mere familiarity of a stock (Oster & Koesterich, 2013), doctors can make diagnostic errors due to patients' disruptive behaviors (Schmidt et al., 2016), or judges can misinterpret evidence because of intuitive stereotypical associations (Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006; Thompson & Schumann, 1987). Reasoning and decision-making studies often attribute this bias to the human tendency to base judgments on fast intuitive impressions rather than on more deliberate reasoning (e.g., Evans, 2008; Kahneman, 2011; Stanovich & West, 2000; Thompson, Prowse Turner, & Pennycook, 2011). This intuitive or so-called "heuristic" thinking can be useful because it is fast and effortless and frequently cues good decisions. However, the problem is that heuristics can also cue decisions that conflict with more logical considerations. In this case, following the intuitive heuristic will lead you astray.

Researchers have been studying heuristic bias empirically with reasoning problems in which an intuitively cued heuristic response conflicts with elementary logical principles. One task that has been studied extensively in the reasoning and decision literature is the bat-and-ball problem in Frederick's (2005) Cognitive Reflection Test (CRT), which states:

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

This work was supported by the Netherlands Organization for Scientific Research (project number 409-15-203) and by the Agence Nationale de la Recherche, France (DIAGNOR, ANR–16–CE28–0010–01).

Address correspondence to Eva Janssen e.m.janssen@uu.nl.

Most reasoners intuitively conclude that the ball must cost 10 cents ($1 + $0.10 = $1.10). However, this conclusion is incorrect because in this scenario the bat costs 90 cents more than the ball instead of $1. After some reflection, it should become clear that the correct answer requires a different calculation leading to the conclusion that the ball costs 5 cents ($1.05 + $0.05 = $1.10)[1]. Although the correct "5 cents" answer does not require strong mathematical skills, numerous studies have shown that even educated reasoners fail to solve the problem correctly (Frederick, 2005; Toplak et al., 2014; Travers et al., 2016), even after repeated problem presentation (Meyer et al., 2018; Raoelison & De Neys, 2019; Stagnaro et al., 2018). Correct solution rates among university students are typically very low (about 32%; for a meta-analysis, see Brañas-Garza et al., 2019).

The key goal of the present study was straightforward. We wanted to look at the impact of a simple but somewhat neglected manipulation – the impact of minimal response feedback (i.e., telling participants whether their response is correct or incorrect) – on people's bat-and-ball performance. With minimal response feedback we refer to a simple correct/incorrect assessment that is presented to the reasoner after (s)he has answered a problem. Although there are some isolated exceptions (e.g., Ball, 2013; Zizzo, Stolarz-Fantino, Wen, & Fantino, 2000) such feedback is usually not presented in research on reasoning and heuristic bias (Ball, 2013; Evans, 2002; Kahneman, 2011). Hence, participants are typically not told whether their response is correct or not. In other fields – such as perception and memory research – presenting performance feedback is a common procedure and has sometimes been shown to boost performance (e.g., Ball, Hoyle, & Towse, 2010; Chun & Wolfe, 1996; Donnelly et al., 2007; Hays, Kornell, & Bjork, 2010). Therefore, it might be worthwhile to examine its impact on heuristic bias as well.

In addition to examining whether accuracy feedback can help to improve reasoning performance, it is also relevant to explore how it works. To gain more insight in the nature of the potential effect of feedback, we measured its effects using a two-response paradigm. In the two-response paradigm reasoners first have to enter their initial, intuitive response to a problem and, thereafter, they can take as much time as they need to deliberate and reflect on their final response (Thompson, Prowse Turner, & Pennycook, 2011). To make maximally sure that people do not deliberate during the initial response generation, they have to provide it under time pressure, while – at the same time – their working memory is loaded with a memorization task (Bago & De Neys, 2017, 2019; De Neys, 2017; Newman et al., 2017; Raoelison & De Neys, 2019). Hence, the procedure deprives participants of the resources they need to efficiently deliberate in the initial response stage (Bago & De Neys, 2019). In the current study, we used the two-response paradigm to see whether feedback affected people's intuitive and deliberate reasoning differently. For example, a feedback effect on both initial and final responses would indicate that reasoners can automatize the correct reasoning, whereas an effect on merely final responses would imply that reasoners have to keep correcting their intuitive responses to be able to reason correctly.

We conducted two studies in which we explored the impact of feedback on bat-and-ball reasoning while applying the two-response paradigm. Participants solved three consecutive blocks of problems. Half of the participants received accuracy feedback in the intermediate block. In Study 1, we tested the potential impact of feedback with a multiple-choice response format and a quasi-experimental design. That is, we compared data from a previous study (Raoelison & De Neys, 2019) where participants solved bat-and-ball problems without receiving feedback (no-feedback condition) with new data specifically collected for the purpose of this study, in which bat-and-ball reasoners received the feedback manipulation in the intermediate block (feedback condition). With Study 2, we aimed to replicate the results of Study 1, while applying a full experimental design with a free-response format.

[1] The algebraic equation behind the problem is
$(x + 1) + x = 1.10$.
$2x + 1 = 1.10$
$2x = 0.10$
$x = 0.05$
Hence, the required calculation to solve the problem is $(1.10 - 1.00) / 2 = 0.05$.

In addition to response accuracy, we also logged participants' response times for explorative analyses. First, we wanted to explore whether the provided feedback affected participants' response latencies. Second, we wanted to use the response latencies as a proxy of conflict or error detection. That is, we were interested in whether biased reasoners' showed some sensitivity to the fact that their heuristic response conflicted with the correct response. Such conflict or error detection has been shown to result in an increase in response latencies (when contrasted with latencies for easy control problems in which the heuristic response is also correct, see e.g., Bago & De Neys, 2017, 2019; De Neys, 2012; Frey et al., 2017; Johnson, Tubau, & De Neys, 2016; Pennycook, Fugelsang, & Koehler, 2015). We wanted to explore whether individuals whose accuracy improved after feedback were characterized by differential conflict detection effects.

Study 1

Method

Participants. A total of 50 participants were recruited for the feedback condition (all gave informed consent). Half of the participants (n = 25) were volunteers, recruited via personal networks. The other half (n = 25) was recruited on Prolific Academic (www.prolific.ac) and were paid at a rate of £5/h. All participants were native English speakers (30 females; age: M = 27.4 years, SD = 8.5)[2]. Most participants reported high school (50.0%) or a Bachelor degree (38.0%) as highest completed level of education, followed by a Master's degree (6.0%), less than high school (2.0%) and a Doctoral degree (2.0%), respectively.

We used data of a previous a study (Raoelison & De Neys, 2019) as a base-line to compare with our feedback data. In this study, 62 participants solved the same bat-and-ball problems without receiving feedback (no-feedback condition). All participants in this study were also native English speakers and recruited on Prolific Academic and were also paid at a rate of £5/h (38 females; age: M = 35.5 years, SD = 13.2). Most participants reported a Bachelor degree (46.8%) or high school (35.5%) as highest completed level of education, followed by a Master's degree (12.9%), less than high school (3.2%), and a Doctoral degree (1.6%), respectively.

Materials. The materials for the feedback condition were taken from the study by Raoelison and De Neys (2019), who designed a total of 110 items to test the robustness of biased responding on bat-and-ball problems by examining how responding is affected by repeated problem presentation. We used 33 items out of their 110 items. Of those 33 items, 15 items were variations of the bat-and-ball problem that had the same underlying structure as the original problem but different superficial item content (e.g., "In a company there are 150 men and women in total. There are 100 more men than women. How many women are there?"). Each problem specified two types of objects with different quantities instead of prices (e.g., see Bago & De Neys, 2019; Mata, Ferreira, Voss, & Kollei, 2017). Each of the 15 problems featured unique content with a total amount that was a multiple of ten and ranged from 110 to 650 (Raoelison & De Neys, 2019). Each problem was presented with four answer options; the correct response ("5 cents" in the original bat-and-ball), the intuitively cued "heuristic" response ("10 cents" in the original bat-and-ball), and two foil options. Mathematically speaking, the correct equation to solve the standard bat-and-ball problem is: $1.00 + 2x = $1.10 (see footnote 1), instead, people are thought to be intuitively using the "$1.00 + x = $1.10" equation to determine their response (Kahneman, 2011). We always used the latter equation to determine the "heuristic" answer option, and the former to determine the correct answer option for each problem. Following Bago and De Neys (2019), the two foil options were always the sum of the correct and heuristic answer (e.g., "15 cents" in original bat-and-ball units) and their second greatest common divider (e.g., "1 cent" in original units). For each item, the four response options appeared in a randomly determined order. The following illustrates the full item format:

[2] Demographic information of one participant is missing because he/she dropped out before completing the demographic questions at the end.

In a company there are 150 men and women in total.
There are 100 more men than women.
How many women are there?
- o 50
- o 75
- o 5
- o 25

One possible cause for a lack of an intervention effect is that participants simply become bored with the repeated problem presentation and stop paying attention. To avoid that the task would become too repetitive and to verify that participants stayed minimally engaged in the task there were also 15 control problems. In the standard bat-and-ball versions the intuitively cued "heuristic" response cues an answer that conflicts with the correct answer, hereafter referred to as "conflict" problems. In the "no-conflict" control problems, the heuristic intuition was made to cue the correct response option. This was achieved by deleting the critical relational "more than" statement (De Neys et al., 2013; Travers et al., 2016). With the above example, a control problem version would look as follows:

In a company there are 150 men and women in total.
There are 100 men.
How many women are there in the company?
- o 50
- o 75
- o 5
- o 25

In this case the intuitively cued "50" answer was also correct. We presented the same four answer options as for a corresponding standard conflict version. We added three words to the control problem question (e.g., "in the company") so that standard "conflict" and control "no-conflict" versions had roughly the same length. Given that the control items can be solved correctly on the basis of mere intuitive reasoning, we expected to see ceiled performance on the control items throughout, if participants are paying minimal attention to the task and refrain from mere random responding.

Finally, in addition to the 15 conflict and the 15 no-conflict problems, there were also 3 filler problems in which participants simply had to add two quantities. For example,

A tech company is offering 100 Motorola phones and 10 Samsung phones.
How many phones are they offering in total?
- o 110
- o 90
- o 250
- o 1000

The rationale behind the filler problems was that these would further help to render the task less repetitive and predictable. In total, participants had to solve 33 problems. The problems were grouped into three blocks (i.e., pretest, intermediate, and posttest) containing each 5 standard conflict problems, 5 control no-conflict problems, and one filler problem. The filler problem was always presented as the sixth problem in a block. Conflict and no-conflict problems were presented in a randomized order. Participants could take a short break after completing each block. We logged both response accuracy and response times on all 33 problems.

As noted, data from the original Raoelison and De Neys (2019), was used as no-feedback baseline. Blocks were similarly structured as the blocks in the feedback condition. In the Raoelison and De Neys study, participants solved a total of 10 blocks. Only the data from the first 3 blocks were analyzed in the current study[3].

Procedure. The experiment was run online on the Qualtrics platform. Participants were specifically instructed that the experiment demanded their full attention throughout. As in Raoelison & De Neys (2019), the study adopted the two-response procedure from Bago and De Neys (2017, 2019). Participants were instructed they had to provide two consecutive responses for each problem. They were told that we were interested in their very first, initial answer that came to mind and that – after selecting their initial response – they could reflect on the problem and take as much time as they needed to provide a final answer. To minimize the possibility that participants deliberated during the initial response stage, the initial response had to be generated within a stringent response deadline and while cognitive resources were burdened with a secondary load task. The deadline for the initial response was set to 5 s, based on the pretesting of Bago and De Neys (2019) who established that this amounted to the time needed to read the problem. The load task was based on the dot memorization task (Miyake et al., 2001). Before each reasoning problem, participants were presented with a complex visual pattern (i.e., 4 crosses in a 3×3 grid) they had to memorize while solving the reasoning problem. After answering the reasoning problem the first time (intuitively), participants were shown four different matrices and had to choose the correct, to-be-memorized pattern (see Raoelison & De Neys, 2019). The load and deadline were applied only during the initial response stage and not during the subsequent final response stage in which participants were allowed to deliberate.

Every trial started with a fixation cross shown for 2000 ms. We then presented the first sentence of the problem (e.g., "In a company there are 150 men and women in total.") for 2000 ms. Next, the target pattern for the memorization task was presented for 2000 ms. Afterwards the full problem was presented. At this point, participants had 5000 ms to give an answer; after 4000 ms the background of the screen turned yellow to warn participants about the upcoming deadline. If they did not provide an answer before the deadline, they were asked to pay attention to provide an answer within the deadline on subsequent trials. After the initial response was entered, participants were presented with four matrix patterns from which they had to choose the correct, to-be-memorized pattern. Once they provided their memorization answer, they received feedback as to whether it was correct. If the answer was not correct, they were also asked to pay more attention to memorizing the correct pattern on subsequent trials.

Finally, the same item was presented again, and participants were asked to provide a final response. The presentation order of the response options was always the same in the initial and final response stage but was randomized across trials. Once participants clicked on one of the answer options they were automatically advanced to the next trial. The color of the answer options was green during the first response, and blue during the final response phase, to visually remind participants of which question they were answering. Therefore, right under the question we also presented a reminder sentence: "Please indicate your very first, intuitive answer." and "Please give your final answer.", respectively, which was also colored as the answer options. At the very end of the experiment, participants were shown the standard bat-and-ball problem and were asked whether they had seen it before. We also asked them to enter the solution. Finally, participants completed a page with demographic questions.

Feedback manipulation. In both the feedback and no-feedback condition participants were informed when they had completed a block (i.e., after 11 trials), how many blocks were still left and instructed to press the 'Next' button when they were ready to continue with the next block. Participants in the feedback condition

[3] Because the 10 blocks in Raoelison and De Neys (2019) were presented in a randomized order and the current study only used participants' first 3 blocks, the exact superficial item content was a random selection of the 110 items in the 10 blocks and hence differed in both conditions.

additionally received the feedback manipulation in the second (intermediate) block. After completing the pretest block, they received the standard information message that they had completed a block and were additionally informed that they would receive feedback about their performance at the intuitive stage in the next block. The feedback was given immediately after participants had entered their intuitive response. The feedback said either "Correct!" or "Incorrect!". Participants then had to click on 'Next' to complete the load task and to give their final response. The feedback was given after all 11 problems during this block.

Exclusion criteria. We analyzed all conflict and no-conflict trials, which were 30 trials × 112 participants = 3360 trials in total. Participants failed to provide their first answer before the deadline on 113 trials (3.4% of all trials) and further failed to pick the correct matrix for the load task on 353 trials (10.9% of remaining trials). Since we could not guarantee that the initial response for these trials did not involve any deliberation, we discarded them and analyzed the 2894 remaining trials (86.1% of all trials). On average each participant contributed 12.9 (*SD* = 2.0) conflict trials and 13.0 (*SD* = 2.1) no-conflict trials. Note that following Raoelison & De Neys (2019) we did not exclude participants based on their response to the familiarity question[4].

Results and Discussion

Response accuracy. For each participant, we calculated the average proportion of correct initial and final responses on the conflict problems and no-conflict problems in each of the three blocks (pretest, intermediate, posttest). Figure 1 (top panel) provides an overview of the average performance on the conflict problems of the feedback and the no-feedback condition. We first focus on the response accuracies for the final responses. Figure 1 shows that most participants failed to solve the conflict problems correctly in the first block (pretest). On average, the final response accuracy at the pretest was 28.2% (*SE* = 6.0) for the feedback condition and 27.7% (*SE* = 5.4) for the no-feedback condition. Both conditions improved in average performance from pretest to posttest, but the feedback condition improved, with an average increase of 14.9% (*SE* = 4.5), more than the no-feedback condition, which had an average increase of 9.4% (*SE* = 3.0). An ANOVA on final accuracy with block (pretest vs. posttest) as within-subjects factor and condition (feedback vs. no feedback) as between-subjects factor revealed a main effect of block, $F(1, 110) = 20.77$, $p < .001$, $\eta^2_p = .159$. There was, however, no main effect of condition, $F(1, 110) = 0.16$, $p = .688$, $\eta^2_p = .014$, neither an interaction of block with condition, $F(1, 110) = 1.12$, $p = .293$, $\eta^2_p = .010$. Thus, reasoners in both conditions improved their final accuracy on the conflict reasoning problems over time, but – despite a small observed trend towards a better improvement for the feedback condition – there was no effect of feedback on participants' reasoning performance.

We repeated all the analyses on the final accuracies with the initial accuracies. Except for slightly lower performance averages, the results were fully consistent (see also Figure 1). On average, the initial response accuracy at the pretest was 18.6% (*SE* = 4.4) for the feedback condition and 23.0% (*SE* = 4.4) for the no-feedback condition. Both conditions improved in average performance from pretest to posttest, but the feedback condition improved, with an average increase of 23.0% (*SE* = 5.4), more than the no-feedback condition, which had an average increase of 11.7% (*SE* = 3.5). Again, the ANOVA showed that reasoners in both conditions improved their accuracy from pretest to posttest but were not affected by the feedback manipulation, block: $F(1, 110) = 30.60$, $p < .001$, $\eta^2_p = .218$, condition: $F(1, 110) = 05$, $p = .828$, $\eta^2_p = .002$; block×condition interaction: $F(1, 110) = 3.68$, $p = .058$, $\eta^2_p = .032$.

As expected, for the no-conflict control problems we observed a performance at ceiling for both conditions in all blocks with grand means of 95.8% (*SE* = 0.7) and 99.0% (*SE* = 0.3) for initial and final accuracy, respectively (see also Figure S1 in the Supplementary Material). An ANOVA on initial accuracy showed that there was no main effect of block, $F(1, 110) = 3.35$, $p = .070$, $\eta^2_p = .030$, no effect of condition, $F(1, 110) = 0.65$, $p =$

[4] Exploratory analyses confirmed that the pattern of results was similar when this exclusion criterion was applied or not. Reported results concern the full dataset.

.422, η^2_p = .006, or an interaction, $F(1, 110) = 0.76$, p = .386, η^2_p = .007. For final accuracies none of the factors reached significance, block: $F(1, 110) = 0.96$, p = .331, η^2_p= .009; condition: $F(1, 110) = 0.21$, p = .646, η^2_p = .002; block×condition: $F(1, 110) = 0.46$, p = .833, η^2_p < .001.

In sum, as observed by Raoelison and De Neys (2019), mere repeated problem presentation resulted in a slight performance increase on standard bat-and-ball problems. However, providing feedback did not result in a further significant improvement per se[5].

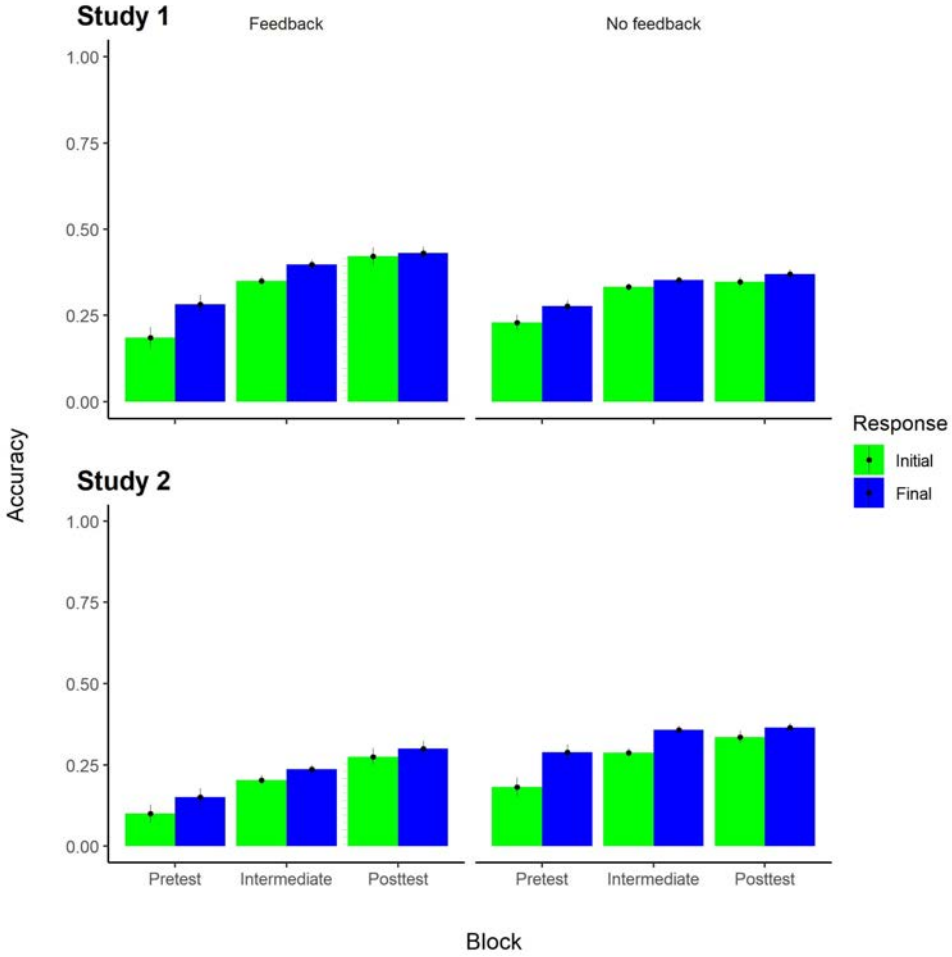

Figure 1. Average initial and final accuracy on conflict problems. Error bars are standard errors.

Individual level directions of change. To gain more insight into how participants solved the problems, we additionally performed a direction of change analysis (Bago & De Neys, 2017, 2019) for each individual participant. More specifically, on each trial people can give a correct or incorrect response in each of the two response stages. Consequently, this can result in four different types of answer patterns on any single trial ("00", incorrect response in both stages; "11", correct response in both stages; "01", initial incorrect and final correct response; "10", initial correct and final incorrect response). Figure 2 (top panel) plots the direction of change

[5] To be absolutely sure that the feedback had no effect, we conducted some additional explorative analyses:
(1) we conducted mixed effect logistic regression analyses on the dichotomous item accuracy responses with subjects' ID as random effect, yielding comparable results (see Table S1 in the Supplementary Material); (2) we controlled for effects of sex as covariate or moderator (Bosch-Domènech et al., 2014; Brañas-Garza et al., 2019), also yielding similar results (see Tables S2-S4 in the Supplementary Materials); and (3) we repeated all the analyses on a subsample in which we only included the participants that failed to solve any of the pretest conflict problems (i.e., the participants that already solved pretest conflict problems correctly could not further benefit from the feedback), again yielding similar results.

classification on each of the consecutive 15 conflict problems (5 per block) for each individual participant. As the figure indicates, the overall pattern was very similar in the feedback and no-feedback condition. We first describe the main trends applying to both conditions and end with a comparison between them. By and large, as in Raoelison and De Neys (2019), we can classify the participants in three main groups. First, most participants (66 out of 112 participants or 58.9%) predominantly gave 00 responses from start to finish. Hence, the majority of the participants consistently gave incorrect intuitive and deliberate responses and remained biased throughout the study. This group is labeled as the "biased" group in Figure 2. Second, 22.3% of the participants (25 out of 112) already gave a correct (final) response at their very first trial and predominantly remained responding correctly throughout the study. This group of "correct" reasoners obviously did not need any intervention to arrive at the correct answer. Third, 18.8% (21 out of 112) started with an incorrect (final) response and found the correct answer somewhere along the way. Once they had found the solution, they remained correct on almost all subsequent trials. We labeled this group as the "insight" group. Interestingly, both in the "correct" and "insight" group we see that for most reasoners, correct responding initially occurs during the deliberation (final response) stage (i.e., a "01" response). However, after only one or two trials, they also managed to solve the subsequent problems correctly without deliberation (i.e., already correct at the intuitive response stage, i.e., a "11" response). This suggests that most participants who found the correct answer automatized the reasoning very quickly.

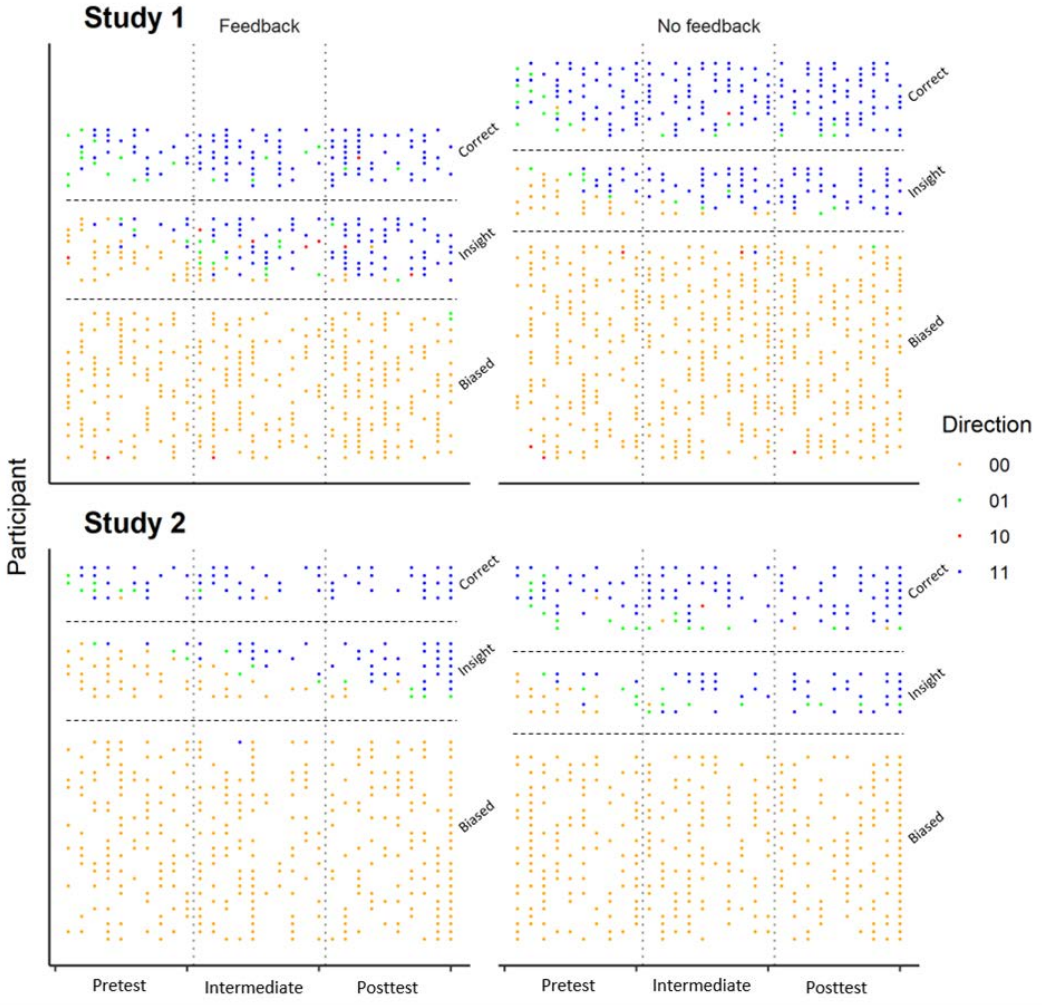

Figure 2. Individual trajectories (each row represents one participant). Due to discarding of missed deadline and load trials (see Exclusion criteria), not all participants contributed 15 analyzable trials. Participants are ranked based on the sum of their total initial and final response accuracy.

We now turn to the comparison between conditions. The key question is whether it was more likely to gain "insight" after receiving accuracy feedback. Hence, we simply tallied how many individuals in the insight group started responding correctly after the onset of feedback (i.e., after the pretest). This was the case for 6 participants (12.0% of total n = 50). For comparison, in the no-feedback condition there were 3 participants (4.8% of total n = 62) who showed the insight pattern after the pretest block. Hence, these effects put the group level effects further in perspective. There is some evidence for a small intervention trend but this trend is driven by only a handful of participants. The vast majority of biased reasoners' is completely unaffected by feedback (or mere repeated presentation in the no-feedback condition).

Response latencies. We additionally explored whether feedback affected participants' response latencies. For each participant, we calculated the average final response time[6] on the conflict problems and on the no-conflict problems in each of the three blocks, while distinguishing between correct and incorrect final responses. To reduce the impact of outliers, we used log-transformations which were then back transformed to enhance the interpretation. Furthermore, we excluded all trials with final response times > 120 s (i.e., > 10.5 *SD*s above average). This concerned two conflict trials and one no-conflict trial. Given the explorative nature, we only report descriptive trends.

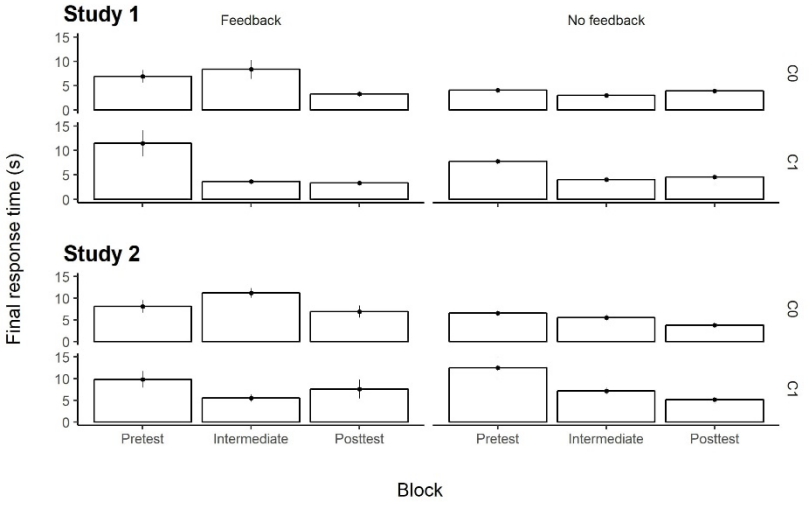

Figure 3. Average final response times on conflict problems. C0 = incorrect performance, C1 = correct performance. Error bars are standard errors.

Figure 3 (top panel) provides an overview of the average final response times of the feedback and the no-feedback condition on the conflict problems (see Figure S2 in the Supplementary Material for an overview of response time on the no-conflict problems). In the no-feedback condition we observed a general speeding-up after the pretest block, both for correct responses (pretest: M = 7.75 s, SE = 1.19; intermediate: M = 4.01 s, SE = 0.51; posttest: M = 4.54 s, SE = 1.19) and incorrect responses (pretest: M = 4.06 s, SE = 0.50; intermediate: M = 3.04 s, SE = 0.32; posttest: M = 3.93 s, SE = 1.26). Hence, people responded faster in the blocks following the pretest block, although the absolute difference with the pretest was minimal for the incorrect responses). We observed the same trend for correct responses in the feedback condition (pretest: M = 11.45 s, SE = 2.67; intermediate: M = 3.67 s, SE = 0.33; posttest: M = 3.34 s, SE = 0.35). However, the incorrect responses in the feedback condition showed a divergent pattern. Here, the average response time increased during the intermediate block (M = 8.37 s,

[6] Given that initial responses were given under fixed response deadline we focus exclusively on the unrestricted final response times.

SE = 1.96), as compared to the pretest (*M* = 6.92 s, *SE* = 1.34) or posttest (*M* = 3.29 s, *SE* = 0.54). This indicates that the (negative) feedback was processed and made people take more time to respond. However, this additional deliberation time did not help (most participants) to arrive at the correct response (see previous section on response accuracy).

Conflict detection. In further exploratory analyses we used participants' final response latencies as proxy of conflict detection[7]. As noted in the introduction, previous studies have suggested that biased reasoners often show some minimal error or conflict sensitivity (Bago & De Neys, 2017, 2019; De Neys, 2012; Frey et al., 2018; Johnson et al., 2016; Pennycook et al., 2015). These studies contrast people's processing of conflict and no-conflict problems. On the no-conflict problems, the intuitively cued heuristic is also correct. On the conflict problems, the intuitively cued heuristic response conflicts with the correct response. If people are sensitive to this conflict, this should affect their processing (e.g., response time). Results indeed show that response times on incorrectly solved conflict problems are typically longer than the response times for correctly solved no-conflict problems (Bago & De Neys, 2017, 2019; De Neys, 2012; Frey et al., 2018; Johnson et al., 2016; Pennycook et al., 2015).

We first checked whether we replicated the results from previous studies. Overall, across both our conditions we found that during the pretest incorrect conflict response (*M* = 5.32 s *SE* = 0.66) indeed took longer than the correct no-conflict response (*M* = 3.85, *SE* = 0.20; i.e., on average a 1.47 s increase). The majority of biased reasoners showed this effect (*n* = 60 out of 89, 67.4%). Next, analogous to the analysis of overall response times, we contrasted the conflict detection effects for each individual (i.e., mean response time incorrect conflict – mean response time correct no-conflict) in the three consecutive blocks, per condition. Figure 4 (top panel) gives a complete overview. As with the overall latency effect, feedback clearly boosted the conflict detection effect during the intermediate block. This suggests that the increased deliberation time under feedback was specifically tied to the presence of conflict.

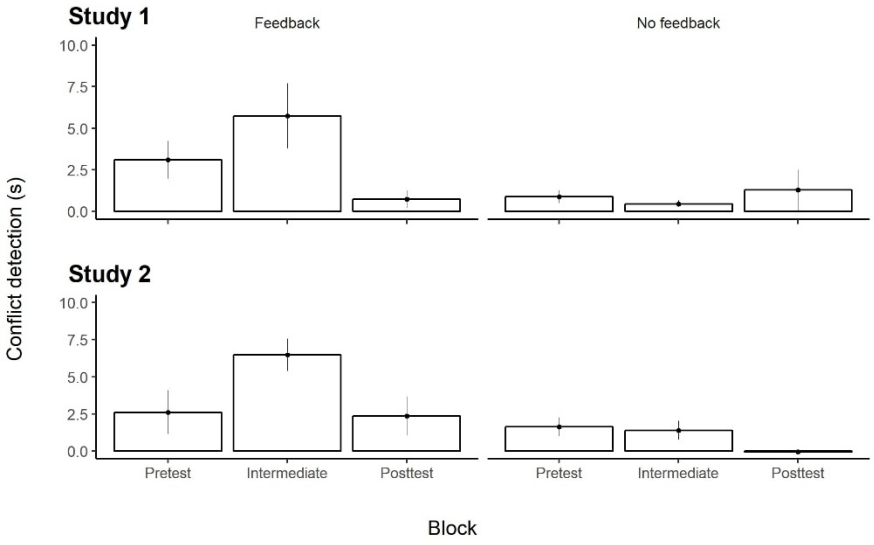

Figure 4. Average conflict detection effect size as indexed by response time. Error bars are standard errors. *Note*. Due to a technical failure, final response time of one no-conflict trial in the intermediate block and one no-conflict trial in posttest block is missing for all participants in Study 2.

[7] Two-response studies have suggested that initial response latencies might not reliably track conflict detection effects at the initial response stage (Bago & De Neys, 2017; Thompson & Johnson, 2014).

Finally, we also wanted to explore whether one's conflict detection was predictive of the intervention effect. That is, we examined whether reasoners whose accuracy improved from pretest to posttest already showed a stronger conflict detection effect in the first pretest block than those who had not improved (see Table 2, top panel). This indeed seemed to be case, in both the feedback and no-feedback conditions, the improved reasoning group had a relatively larger conflict detection effect size average at the pretest when compared to the unimproved reasoning group.

Table 1: Overview of average pretest conflict detection effect size as indexed by response time (s) for improved versus unimproved biased reasoners

	n	*M*	*SE*
Study 1			
Feedback condition			
Improved reasoners	13	5.55	2.83
Unimproved reasoners	26	1.88	0.91
No-feedback condition			
Improved reasoners	12	2.54	1.42
Unimproved reasoners	38	0.34	0.20
Study 2			
Feedback condition			
Improved reasoners	9	9.44	5.73
Unimproved reasoners	27	0.74	0.35
No-feedback condition			
Improved reasoners	7	3.15	2.65
Unimproved reasoners	25	1.21	0.36

Note. The *n*s do not add up to the total sample sizes because 23 unimproved participants in Study 1 and 12 unimproved participants in Study 2 had a ceiled pretest performance (i.e., conflict detection effect size can only be calculated for participants who gave at least one biased response).

Study 2

With Study 2, we aimed to test the robustness of the Study 1 findings while applying a number of methodological optimizations. First, we adopted a full experimental design instead of a quasi-experimental design and randomly allocated participants to a feedback and no-feedback condition. Second, we adopted a free-response format instead of multiple-choice answering options, to eliminate the possibility that a potential effect of feedback was driven by reasoning backwards from the presented answering options. Third, after the last problem in the posttest, participants were asked to justify their answer, to see whether they were able to explain the reasoning behind their answer. Fourth, participants had to solve two transfer problems after completing the bat-and-ball problems, to see whether a potential feedback effect also led to transfer.

Method

Given the similarity of both study designs and to avoid repetition, we only describe the aspects of the method that deviated from Study 1.

Pre-registration. The study design, sample size, and hypotheses were preregistered on the Open Science Framework (https://osf.io/b9u8r/). No specific analyses were preregistered.

Participants. For Study 2, 80 participants (all gave informed consent) were recruited on Prolific Academic (www.prolific.ac), were paid at a rate of £5/h, and were randomly assigned to either the feedback condition (n = 40) or the no-feedback condition (n = 40). All participants were native English speakers (49 females; age: M = 34.3 years, SD = 12.6). For most participants the highest completed level of education was high school (45.0%) or a Bachelor degree (35.0%), followed by a Master's degree (8.8%), less than high school (7.5%), and a Doctoral degree (3.8%), respectively.

Materials. In addition to the 33 problems in Study 1, we administered one justification question and two transfer problems.

Justification. After completing their final bat-and-ball problem, participants received the following message (see Bago & De Neys, 2019):

> You have almost completed Block 3. We are interested in the reasoning behind your response to the final question:
>
> In a school there are 130 boys and girls in total.
> There are 100 more boys than girls. How many girls are there ?
>
> Could you please justify, why do you think that your previously entered response is the correct response to the question? Please choose from the presented options below:
> - o I did the math. Please specify how:_____
> - o I guessed
> - o I decided based on intuition/gut feeling
> - o Other, please specify: _____

Two independent raters (first and second author) judged whether the specified justifications indicated a correct, incorrect, or unspecified justification. They were in agreement in 100% of the cases. A justification was coded as correct when the correct calculation was provided (e.g., "130-100=30 / 30/2=15"). A justification was coded as incorrect when it referred to the incorrect/heuristic math (e.g., "The total number is 130 and the boys are 100 more so I subtracted 100 from the total to give me the answer"). Justifications that did not explain the math procedure were coded as "not specified" (e.g., "I DID THE MATH").

Transfer problems. The transfer problems were two CRT-like items. The first one was an adapted version of the "widget" problem from the original CRT (Frederick, 2005):

> If it takes 10 minutes for ten cooks to prepare 10 hamburgers,
> how long would it take for 200 cooks to prepare 200 hamburgers?

The heuristic response here is 200 minutes and the correct response 10 minutes. The second transfer problem was an item derived from Thomson and Oppenheimer (2016):

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

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

Procedure. Study 2 used a free response format, following the same procedure as Bago and De Neys (2019). Both in the initial and final response stage, participants needed to click on a blank field where they had to enter their response, type their answer, and click on a button labelled "Next" to advance. Because typing an answer requires more time than selecting a multiple-choice answer, the response deadline was 8000 ms instead of 5000 ms (see Bago & De Neys, 2019). After 6000 ms the background of the screen turned yellow to warn participants about the upcoming deadline. Participants were instructed to only type numbers (no letters). Each time after a participant had entered an invalid character during the intuitive response stage, (s)he received a reminder message to only enter numbers. All invalid responses were excluded. In the final response stage it was not possible to enter other characters than numbers. To familiarize participants with the two-response procedure, they first solved two unrelated problems with the initial response deadline but yet without the load task. Next they practiced one load task and, thereafter, they solved two problems following the complete two-response procedure.

The order of all 33 problems was randomized in the same way as in Study 1 except for the fact that all participants completed the same final problem, a conflict problem, so that we could ask for their justification.

After the justification question, participants were instructed that they had completed block 3 out of 4 and that the final block consisted of only two problems that were somewhat different from the previous ones and, additionally, that they could think as long as they wanted to solve these. In contrast to Study 1, we did not check whether participants were familiar with the classic bat-and-ball problem (as we did not use this as an exclusion criterion, see Footnote 4).

Feedback manipulation. We made the accuracy feedback during the intermediate block somewhat more salient than in Study 1. That is, for correct answers the feedback said "**CORRECT answer!**" in green, bold, capitalized letters and for incorrect answers "**INCORRECT answer!**" in red, bold, capitalized letters.

Exclusion criteria. We analyzed all conflict and no-conflict trials, which were 30 trials × 80 participants = 2400 trials in total. Participants failed to provide their first answer before the deadline on 37 trials (1.4% of all trials) and further failed to pick the correct matrix for the load task on 269 trials (11.4% of remaining trials). Since we could not guarantee that the initial response for these trials did not involve any deliberation, we discarded them and analyzed the 2094 remaining trials (87.3% out of 2400 trials). On average each participant contributed 13.1 (*SD* = 1.6) conflict trials and 13.1 (*SD* = 1.5) no-conflict trials.

Results and Discussion

In general, the results of Study 2 were highly similar to the results of Study 1. For completeness, we discuss all results but try to be concise where possible.

Response accuracy. Figure 1 (bottom panel) provides an overview of the average performance of the feedback and no-feedback condition on the conflict problems. For the final accuracies, the average performance at the pretest was 15.1% (*SE* = 5.4) for the feedback condition and 28.8% (*SE* = 6.7) for the no-feedback condition. Both conditions improved in average performance from pretest to posttest, but the feedback condition improved, with an average increase of 14.9% (*SE* = 4.9), more than the no-feedback condition, which had an average increase of 7.7% (*SE* = 3.8). The ANOVA indicated that reasoners in both conditions improved their accuracy from pretest to posttest but were not affected by the feedback manipulation, block: $F(1, 78) = 13.25$, $p < .001$, $\eta^2_p = .211$; condition: $F(1, 78) = 1.26$, $p = .265$, $\eta^2_p = .048$; block×condition: $F(1, 78) = 1.35$, $p = .248$, $\eta^2_p = .01$.

As in Study 1, the initial accuracies showed the same pattern as the final accuracies but with slightly lower performance averages (see also Figure 1). The average initial accuracy at the pretest was 10.0% (*SE* = 4.1) for the feedback condition and 18.3% (*SE* = 5.2) for the no-feedback condition. Both conditions improved in

average performance from pretest to posttest, but the feedback condition improved, with an average increase of 17.5% (*SE* = 5.3), more than the no-feedback condition, which had an average increase of 15.3% (*SE* = 4.9). Again, the ANOVA indicated that reasoners in both conditions improved their accuracy from pretest to posttest but were not affected by the feedback manipulation, block: $F(1, 78) = 20.87$, $p < .001$, $\eta^2_p = .243$; condition: $F(1, 78) = 0.83$, $p = .364$, $\eta^2_p = .048$; block×condition: $F(1, 78) = 0.10$, $p = .754$, $\eta^2_p = .007$.

The no-conflict control problems also showed the expected pattern. That is, we observed a performance at ceiling for both conditions in all blocks with grand means of 98.6% (*SE* = 0.4) and 99.7% (*SE* = 0.2) for initial and final accuracy, respectively (see also Figure S1 in the Supplementary Material). An ANOVA on initial accuracy showed that there was a small main effect of block, $F(1, 78) = 7.03$, $p = .010$, $\eta^2_p = .083$, no effect of condition, condition: $F(1, 78) = 0.37$, $p = .548$, $\eta^2_p = .004$, or an interaction, $F(1, 78) = 0.01$, $p = .933$, $\eta^2_p < .001$. For final accuracies none of the factors reached significance, block: $F(1, 78) = < 0.001$, $p > .999$, $\eta^2_p < .001$; condition: $F(1, 78) = < 0.001$, $p = > .999$, $\eta^2_p < .001$; block×condition: $F(1, 78) = 2.00$, $p = .161$, $\eta^2_p = .025$.

Thus, also when using a free-response format (Study 2) instead of multiple-choice options (Study 1), reasoners improved their accuracy on the conflict bat-and-ball problems from pretest to posttest but were not affected by the feedback manipulation per se (despite a small observed trend towards a better improvement for the feedback condition)[8].

Individual level directions of change. Figure 2 (bottom panel) plots the direction of change classification on each of the consecutive conflict problems for each individual participant. Just as in Study 1, the figure shows a very similar pattern in the feedback and no-feedback conditions. First, most participants (53 out of 80 participants or 65.0%) were in the "biased" group. That is, they predominantly gave incorrect intuitive and deliberate (00-) responses and remained biased throughout the study. Second, 17.5% of the participants (14 out of 80) was in the "correct" group, meaning that they already gave a correct (final) response at their very first trial and predominantly remained responding correctly throughout the study. Third, 17.5% (14 out of 80) was in the "insight" group. These participants started with an incorrect response and found the correct answer somewhere along the way, and remained correct from then on. Again, for both the "correct" and "insight" group, we observed that most participants who found the correct answer automatized the reasoning very quickly (i.e., only on or two 01-responses preceded the consistent row of 11-responses).

With regard to the comparison between conditions, 5 participants in the feedback condition (12.5% of total $n = 40$) showed the insight pattern after the pretest block (i.e., after onset of feedback), compared to 1 participant (2.5% of total $n = 40$) in the no-feedback condition. Hence, in line with Study 1, we observed evidence for a small intervention trend but this trend is driven by only a handful of participants. The vast majority of biased reasoners was not affected by feedback (or mere repeated presentation in the no-feedback condition).

Response latencies. Figure 3 (bottom panel) provides an overview of the average final response times of the feedback and no-feedback condition on the conflict problems (see Figure S2 in the Supplementary Material for an overview of response time on the no-conflict problems). For consistency with Study 1, we again excluded all trials with final response times > 120 s. This concerned six conflict trials and two no-conflict trials. We found a similar pattern as in Study 1. In the no-feedback condition we observed a general speeding-up after the pretest block, both for correct responses (pretest: *M* = 12.45 s, *SE* = 2.60; intermediate: *M* = 7.14 s, *SE* = 0.78; posttest: *M* = 5.20 s, *SE* = 0.34) and incorrect responses (pretest: *M* = 6.58 s, *SE* = 0.98; intermediate: *M* = 5.51 s, *SE* = 0.77; posttest: *M* = 3.79 s, *SE* = 0.28). We observed the same trend for correct responses in the feedback condition (pretest: *M* = 9.80 s, *SE* = 1.91; intermediate: *M* = 5.50 s, *SE* = 0.83; posttest: *M* = 7.58 s, *SE* = 2.11) but not for the incorrect responses. The incorrect responses in the intermediate feedback block again showed increased response time (intermediate: *M* = 11.18 s, *SE* = 1.12), as compared to the other two blocks (pretest: *M* = 8.12 s, *SE* = 1.48;

[8] As in Study 1, we conducted some additional (explorative analyses) to be absolutely sure that the feedback had no effect (see Footnote 5). These analyses again yielded no effect of feedback.

posttest: M = 6.93 s, SE = 1.36). This indicates that the (negative) feedback was processed and made people take more time to respond. However, this additional deliberation time did not help (most) participants to arrive at the correct response (see previous section on response accuracy). Again, we found that the improved reasoners had longer response times (n = 6, M = 17.10, SE = 3.40) after receiving negative feedback than those who had not improved (n = 27, M = 9.86, SE = 1.01)[9]. The increase in comparison with the pretest, however, was observed in both groups (improved reasoners: M = 10.75 s, SE = 3.21; unimproved reasoners: M = 3.78 s, SE = 0.85) although it was again smaller for the unimproved reasoners (an overview of the improved reasoners' performance pattern is shown in Figure S3 in the Supplementary Material).

Conflict detection. Consistent with Study 1, we replicated the conflict detection effect as found in previous studies (e.g., Bago & De Neys, 2019; Frey et al., 2017). Across both our conditions we found that during the pretest, incorrect conflict responses (M = 7.39 s SE = 0.91) indeed took longer than the correct no-conflict responses (M = 5.51, SE = 0.29; i.e., on average a 1.88 s increase). The t majority of biased reasoners showed this detection effect (n = 48 out of 68, 70.6%). Figure 4 (bottom panel) gives a complete overview of the conflict detection effects in the three consecutive blocks per condition. As with the overall latency effect and consistent with what we observed in Study 1, feedback clearly boosted the conflict detection effect during the intermediate block. Finally, we again also found that one's conflict detection was predictive of the intervention effect (see Table 1, bottom panel). In both the feedback and no-feedback condition, the improved reasoning group had a relatively larger conflict detection effect size average at the pretest when compared to the unimproved reasoning group.

Justification. The interested reader can find an overview of response justifications analysis in the Supplementary Material Table S5. Results showed that almost all participants that had solved the last conflict item correctly also gave the correct math justification (feedback condition: 11 out of 13; no-feedback condition: 14 out of 14). The majority of the participants that solved the item incorrectly, gave an incorrect math justification (feedback condition: 16 out of 24; no-feedback condition: 18 out of 25). Hence, irrespective of feedback, people who responded correctly typically also managed to explicate the correct solution strategy.

Transfer problems. For each participant, we calculated the average proportion of correct responses on the two transfer problems. Overall, average performance on the transfer problems was more or less similar in the feedback (M = 46.3%, SE = 6.6%) and no-feedback condition (M = 52.5%, SE = 6.7%). We also explored whether the transfer problem performance differed for improved versus unimproved reasoners. As Table S6 in Supplementary Materials shows, this was indeed the case. Both in the feedback and no-feedback condition, an improved reasoner (n = 16, M = 59.4%, SE = 9.4%) was more likely to solve the transfer problems correctly than an unimproved reasoner (n = 64, M = 46.88%, SE = 5.3%)[10].

General Discussion

In the present two studies we tested the impact of a minimal intervention – response accuracy feedback – on people's bat-and-ball performance. We presented participants 15 standard (conflict) and 15 control (no-conflict) versions of the bat-and-ball problem, in three consecutive blocks (pretest, intermediate, and posttest). Half of the participants received accuracy feedback during the intermediate block, whereas the other half did not. Overall, the results of both studies were very consistent and clearly indicated that feedback had, on average, no significant effect on participants' bat-and-ball accuracy. We only observed a small trend in a handful of participants. Our explorative analyses did reveal a trend towards a feedback effect on response latencies (i.e.,

[9] A total of 3 improved participants and 4 unimproved participants in the feedback condition did not enter any incorrect responses in the intermediate (feedback) block and were thus not included in this analysis.

[10] We also ran a control analysis in which reasoners who were at ceiling in the pretest (n = 12) were excluded from the unimproved group. Results led to the same conclusion (see Table S6).

longer response times after receiving negative feedback) and conflict detection (i.e., larger conflict detection effect after receiving negative feedback). Hence, it seemed that feedback evoked extra deliberation but did not help most participants to arrive at the correct response. Interestingly, the small group of reasoners who did learn to correct their errors after receiving feedback, showed a stronger conflict detection effect in the pretest block, took more deliberation time after receiving negative feedback, and performed better on the two transfer problems (in Study 2), compared to the group of reasoners that remained biased on all problems.

Why was the feedback not more effective for improving bat-and-ball accuracy? One suggestion is related to the nature of error on this task. Previous studies (and the current study) revealed that the majority of the biased bat-and-ball reasoners already show some minimal error or bias detection from the onset (Bago & De Neys, 2019; Frey et al., 2017; Gangemi, Bourgeois-Gironde, & Mancini, 2015; Hoover & Healy, 2019; Mata, 2019; but see also Mata et al., 2017; Mata, Schubert, & Ferreira, 2014; Travers et al., 2016). Hence, even without feedback people at least seem to implicitly detect that their answer is not fully warranted. In this light, it is perhaps not surprising that telling them this explicitly has little effect. In other words, people are not biased because they do not realize that 10 cents is incorrect but rather because they do not know how to arrive at the correct solution strategy. Given that most people are capable of solving the algebra behind the bat-and-ball problem (Hoover & Healy, 2017), it remains an open question why they do not arrive at the correct solution strategy themselves. One possible explanation is that the required algebraic solution strategy concerns a reasoning strategy that is not frequently used in the daily-life reasoning of most people and is thus less easily available in long-term memory. Hence, a much stronger or more informative retrieval cue would be needed to arrive at the algebraic strategy. In this sense, giving people more detailed feedback or tutoring about the correct solution strategy might prove more effective.

Nevertheless, in addition to the large majority that remained biased, our results (and Raoelison & De Neys, 2019) also show that some individual participants do manage to arrive at the correct solution strategy themselves. Here, we can distinguish between participants that were right from the start (the "correct" group) and participants that managed to correct themselves after responding biased at first (the "insight" group). For the "insight" group, we believe that our results on the predictive conflict detection effects are especially interesting. Previous studies have suggested there are individual differences in the extent of the conflict signal (Bago et al., 2019; Bago & De Neys, 2017, 2019; Frey et al., 2018; Pennycook et al., 2015). The current results indicated that those reasoners who improved from pretest to posttest (through the feedback or through mere repeated exposure), already showed a more pronounced conflict detection effect (i.e., higher latency increase) before the intervention. Hence, for those with the strongest signal, a minimal intervention "nudge" did suffice to get them to start reasoning correctly. In other words, the people who strongly feel that the 10 cents is incorrect are more likely to arrive at the solution strategy themselves after only a small push to think a little further. Following our previous "availability" suggestion, this group might show stronger conflict detection at the start and could correct themselves because the proper algebraic solution strategy was more easy available in long-term memory than for those who remained biased (e.g., because they have better numeracy skills, greater cognitive ability, or use the skill more often in daily life). This may also explain why they performed better on the transfer tasks than the unimproved "biased" group.

Finally, the "correct" group obviously did not need any intervention to arrive at the correct answer, and was therefore not affected by the feedback manipulation or the repeated problem presentation. We suggest that the solution strategy was most easily available for this correct group, which is in line with previous research on the CRT, showing that those with better numeracy skills, greater cognitive capacity, or work in a domain that requires numeracy skills are more likely to perform correctly (Campitelli & Gerrans, 2014; Janssen et al., 2019; Toplak et al., 2011, 2014).

Obviously, the limited average impact of our feedback intervention does not entail that it is impossible to tutor people on the bat-and-ball problem. Our goal in the present study was to focus on a simple intervention that has proven to be effective in other fields. As we alluded to above, one could try to boost efficacity by

switching to a more extensive tutoring or training in which participants are informed about how to arrive at the correct solution strategy. For example, Hoover and Healey (2017) showed that properly instructing people about the underlying mathematical equation can help to boost performance in the bat-and-ball problem. In this light, it might be interesting to explore the impact of reasoning feedback that may guide participants' reasoning towards the required calculation (e.g., explaining and highlighting the role of the "more than" phrase, adding the correct equation, etc.). Furthermore, although the simple response feedback was not an effective de-biasing strategy for the bat-and-ball problem, it would be interesting to explore its impact on other problems, for example on the classic base-rate problem (De Neys et al., 2011). In this problem the correct solution strategy is based on a very elementary statistical rule which is easy available in long-term memory of most people. If our suggestion concerning the ease of availability of a solution strategy is true, then accuracy feedback could have a stronger de-biasing impact here.

Finally, the current research had several limitations. First, the observed trends in both studies suggest that there might actually be a small effect of feedback that we were unable to detect with current sample size (which only allowed for picking up on medium effects). Second, accuracy feedback was given while participants were still burdened with the load task, which might have had an effect on their attention to the feedback. Hence, even though the response latencies suggested that participants processed the negative feedback, it could be that the feedback would have been more effective when presented after the load task. Third, the overall obtained accuracies were rather low in comparison with other studies (Brañas-Garza et al., 2019), which might indicate that the current participants were not motivated enough to improve their reasoning. In this light, (monetary) incentives would perhaps have made the feedback more effective, although a recent meta-analysis indicated that monetary incentives have no overall effect on bat-and-ball performance (Brañas-Garza et al., 2019).

In closing, although the overall observed impact of feedback on people's accuracy was very limited, we believe that the present paper does indicate that a feedback manipulation has potential as a methodological tool and warrants to be further explored in the reasoning and decision making field.

References

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*, 257–299. https://doi.org/10.1080/13546783.2018.1507949

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

Ball, L. J. (2013). Microgenetic evidence for the beneficial effects of feedback and practice on belief bias. *Journal of Cognitive Psychology*, *25*, 183–191. https://doi.org/10.1080/20445911.2013.765856

Ball, L. J., Hoyle, A. M., & Towse, A. S. (2010). The facilitatory effect of negative feedback on the emergence of analogical reasoning abilities. *British Journal of Developmental Psychology*, *28*, 583–602. https://doi.org/10.1348/026151009X461744

Bosch-Domènech, A., Brañas-Garza, P., & Espín, A. M. (2014). Can exposure to prenatal sex hormones (2D:4D) predict cognitive reflection? *Psychoneuroendocrinology*, *43*, 1–10. https://doi.org/10.1016/j.psyneuen.2014.01.023

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2019). Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics*, *82*, 101455. https://doi.org/10.1016/j.socec.2019.101455

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, *42*, 434–447. https://doi.org/10.3758/s13421-013-0367-9

Chun, M. M., & Wolfe, J. M. (1996). Just say No: How are visual searches terminated when there is no target present? *Cognitive Psychology*, *30*, 39–78. https://doi.org/10.1006/cogp.1996.0002

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*, 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (Ed.). (2017). *Dual process theory 2.0*. Routledge, Taylor & Francis Group.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Correction: Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, *6*. https://doi.org/10.1371/annotation/1ebd8050-5513-426f-8399-201773755683

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

Donnelly, N., Cave, K., Greenway, R., Hadwin, J. A., Stevenson, J., & Sonuga-Barke, E. (2007). Visual search in children and adults: Top-down and bottom-up mechanisms. *Quarterly Journal of Experimental Psychology*, *60*, 120–136. https://doi.org/10.1080/17470210600625362

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of black defendants predicts capital-sentencing outcomes. *Psychological Science*, *17*, 383–386. https://doi.org/10.1111/j.1467-9280.2006.01716.x

Evans, J. St. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, *128*, 978–996. https://doi.org/10.1037//0033-2909.128.6.978

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

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 25–42. https://doi.org/10.1257/089533005775196732

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology*, 1–52. https://doi.org/10.1080/17470218.2017.1313283

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—In search of a phenomenon. *Thinking & Reasoning*, *21*, 383–396. https://doi.org/10.1080/13546783.2014.980755

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review*, *17*, 797–801. https://doi.org/10.3758/PBR.17.6.797

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*, 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

Janssen, E. M., Meulendijks, W., Mainhard, T., Verkoeijen, P. P. J. L., Heijltjes, A. E. G., van Peppen, L. M., & van Gog, T. (2019). Identifying characteristics associated with higher education teachers' Cognitive Reflection Test performance and their attitudes towards teaching critical thinking. *Teaching and Teacher Education*, *84*, 139–149. https://doi.org/10.1016/j.tate.2019.05.008

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*. Lane.

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

Mata, A., Schubert, A.-L., & Ferreira, M. B. (2014). The role of language comprehension in reasoning: How "good-enough" representations induce biases. *Cognition*, *133*, 457–463. https://doi.org/10.1016/j.cognition.2014.07.011

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision Making*, *13*, 246–259.

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*, 621–640. https://doi.org/10.1037//0096-3445.130.4.621

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*, 1154–1170. https://doi.org/10.1037/xlm0000372

Oster, N., & Koesterich, R. (2013). Breaking bad behaviors: Understanding investing biases and how to overcome them. *IShares Market Perspectives*, 1–9.

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

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*, 178–178.

Schmidt, H. G., van Gog, T., CE Schuit, S., Van den Berge, K., LA Van Daele, P., Bueving, H., Van der Zee, T., W Van den Broek, W., LCM Van Saase, J., & Mamede, S. (2016). Do patients' disruptive behaviours influence the accuracy of a doctor's diagnosis? A randomised experiment. *BMJ Quality & Safety*, *26*, 19–23. https://doi.org/10.1136/bmjqs-2015-004109

Stagnaro, M., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is Stable Across Time. *Judgment and Decision Making*, *13*, 260–267.

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

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*, 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*, 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Thompson, W. C., & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The proscecutor's fallacy and defense attorney's fallacy. *Law and Human Behavior*, *11*, 167–187. https://doi.org/10.2307/1393631

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the Cognitive Reflection Test. *Judgment and Decision Making*, *11*, 99–113.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*, 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*, 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

Zizzo, D. J., Stolarz-Fantino, S., Wen, J., & Fantino, E. (2000). A violation of the monotonicity axiom: Experimental evidence on the conjunction fallacy. *Journal of Economic Behavior & Organization*, *41*, 263–276. https://doi.org/10.1016/S0167-2681(99)00076-1

Supplementary Material

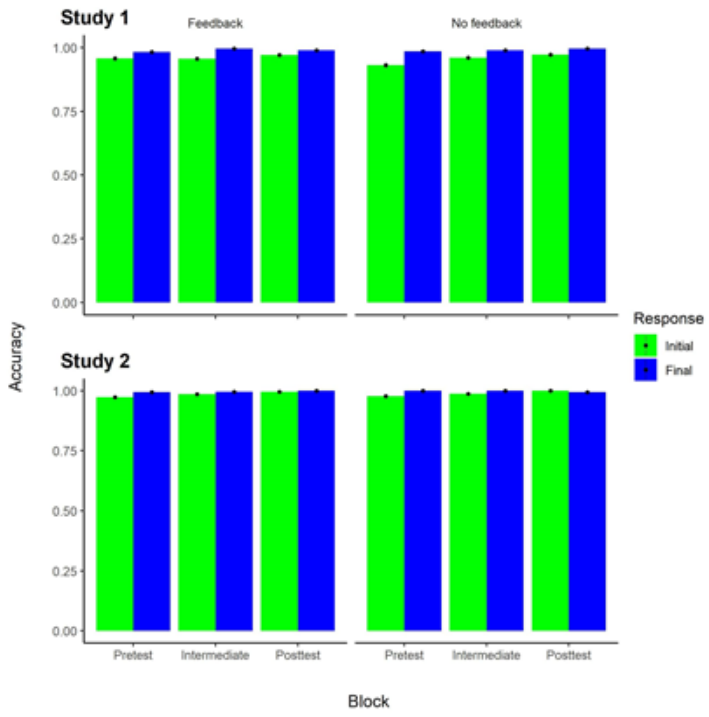

Figure S1. Average initial and final accuracy on no-conflict problems. Error bars are standard errors.

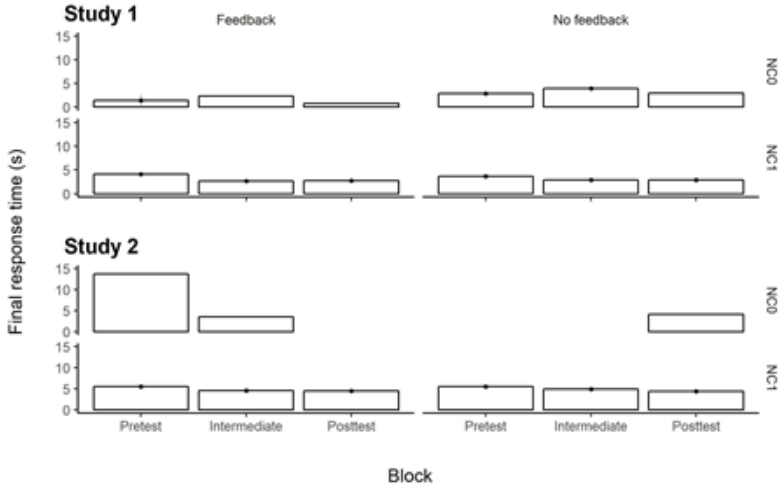

Figure S2. Average, final response times on no-conflict problems. NC0 = incorrect performance, NC1 = correct performance. Error bars are standard errors. *Note*. Due to a technical failure, final response time of one no-conflict trial in the intermediate block and one no-conflict trial in posttest block is missing for all participants in Study 2.

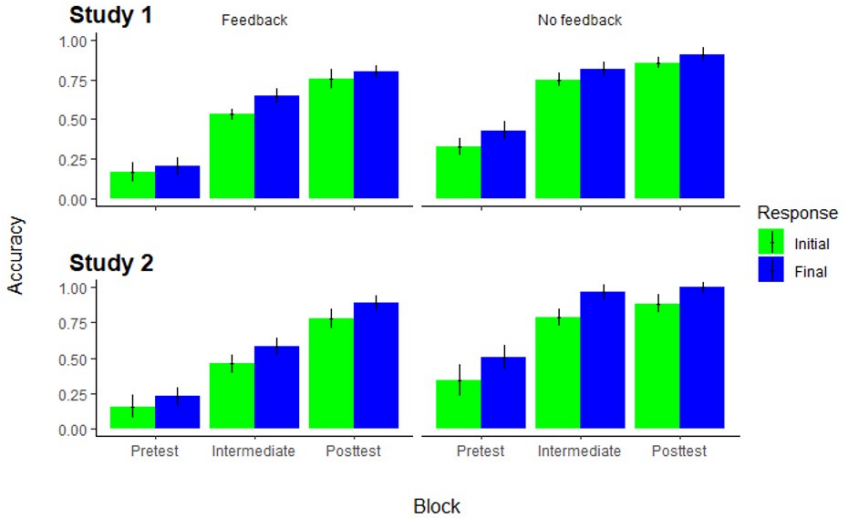

Figure S3. Average initial and final accuracy on conflict problems for improved reasoners only. Error bars are standard errors.

Table S1: Mixed-effects logistic regression model testing the effect of feedback

	Initial responses:	Final responses
Study 1	Coefficient (s.e.)	Coefficient (s.e.)
Fixed effects		
Intercept	-9.19 (1.37)**	-15.39 (2.03)**
Block	2.50 (0.47)**	5.54 (1.49)**
Condition	-0.76 (1.38)	1.56 (2.13)
Block × Condition	1.29 (0.72)	-1.21 (1.65)
Random effects		
Subject	74.81 (8.65)	381.1 (19.52)
Study 2	Coefficient (s.e.)	Coefficient (s.e.)
Fixed effects		
Intercept	-15.88 (2.36)**	-13.44 (1.79)**
Block	6.17 (1.68)**	3.10 (0.86)**
Condition	-4.52 (2.87)	-6.44 (3.77)
Block × Condition	4.44 (2.50)	6.33 (3.42)
Random effects		
Subject	280.6 (16.75)	272.7 (16.51)

Note. Condition coded 0 = no-feedback condition, 1 = feedback condition. Block coded 0 = pretest, 1 = posttest.

* $p < .05$, ** $p < .001$.

Table S2: ANOVAs with sex as covariate and as moderator

	Initial responses				Final responses			
Study 1								
Sex as covariate	*F*	*df*	*p*	η^2_p	*F*	*df*	*p*	η^2_p
Block	30.75	1, 109	< .001	.220	20.83	1, 109	< .001	.160
Condition	0.09	1, 108	.764	.004	0.24	1, 108	.628	.021
Block×Condition	3.92	1, 109	.050	.04	1.22	1, 109	.271	.011
Sex	0.00	1,109	.997	< .001	0.00	1, 108	.966	< .001
Sex as moderator	*F*	*df*	*p*	η^2_p	*F*	*df*	*p*	η^2_p
Block	30.24	1, 107	< .001	.220	20.57	1, 107	< .001	.616
Condition	0.09	1, 107	.764	.004	0.23	1, 107	.629	.021
Block×Condition	3.86	1, 107	.052	.035	1.21	1, 107	.274	.011
Sex	0.00	1, 107	.997	< .001	0.00	1, 107	.966	<. 001
Sex × Block	0.21	1, 107	.649	.002	0.09	1, 107	.767	.001
Sex × Condition	0.23	1, 107	632	.011	0.06	1, 107	.810	.005
Sex × Block × Condition	0.02	1, 107	.880	< .001	0.57	1, 107	.453	.005
Study 2								
Sex as covariate	*F*	*df*	*p*	η^2_p	*F*	*df*	*p*	η^2_p
Block	20.87	1, 78	< .001	.211	13.25	1, 78	<. 001	.120
Condition	0.86	1, 77	.357	.048	1.30	1, 77	.258	.145
Block×Condition	0.10	1, 78	.754	.001	1.35	1, 78	.248	.017
Sex	3.14	1, 77	.080	.157	3.47	1, 77	.071	.261
Sex as moderator	*F*	*df*	*p*	η^2_p	*F*	*df*	*p*	η^2_p
Block	20.46	1, 76	< .001	.212	12.94	1, 76	< .001	.145
Condition	0.85	1, 76	.360	.048	1.29	1, 76.	.260	.120
Block×Condition	0.10	1, 76	.757	.001	1.32	1, 76	.254	.017
Sex	3.12	1, 76	.082	.157	3.33	1, 76	.072	.261
Sex × Block	0.32	1, 76	.575	.004	0.16	1, 76	.694	.002
Sex × Condition	0.29	1, 76	.593	.017	0.62	1, 76	.434	.062
Sex × Block × Condition	0.15	1, 76	.704	.002	0.02	1, 76	.888	< .001

Note. Condition coded 0 = no-feedback condition, 1 = feedback condition. Block coded 0 = pretest, 1 = posttest. Sex coded 0 = male 1 = female.

Table S3: Mixed-effects logistic regression model testing the effect of feedback with sex as covariate

	Initial responses:	Final responses
Study 1	Coefficient (s.e.)	Coefficient (s.e.)
Fixed effects		
Intercept	-9.23 (1.57)**	-15.38 (2.23)**
Block	2.49 (0.47)**	5.54 (1.49(**
Condition	-0.69 (1.41)	1.57 (2.13)
Block × Condition	1.29 (0.72)	-1.21 (1.65)
Sex	0.17 (1.24)	-0.01 (1.46)
Random effects		
Subject	73.3 (8.56)	380.5 (19.51)
Study 2	Coefficient (s.e.)	Coefficient (s.e.)
Fixed effects		
Intercept	-15.19 (2.57)**	-12.75 (2.09)**
Block	6.15 (1.67)**	3.09 (0.86)**
Condition	-4.47 (2.87)	-6.32 (3.78)
Block × Condition	4.38 (2.49)	6.22 (3.42)
Sex	-1.00 (1.71)	-1.05 (1.86)
Random effects		
Subject	272.7 (16.51)	417.1 (20.42)

Note. Condition coded 0 = no-feedback condition, 1 = feedback condition. Block coded 0 = pretest, 1 = posttest. Sex coded 0 = male 1 = female.

* $p < .05$, ** $p < .001$.

Table S4: Mixed-effects logistic regression model testing the effect of feedback with sex as moderator

	Initial responses:	Final responses
Study 1		
Fixed effects	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	-8.91 (1.86)**	-14.65 (2.57)**
Block	2.34 (0.82)**	4.37 (1.79)*
Condition	-0.93 (2.13)	-7.05 (3.65)
Block × Condition	1.08 (1.15)	7.48 (2.96)*
Sex	-0.28 (1.77)	-3.47 (3.96)
Sex × Block	0.21 (0.97)	3.33 (3.47)
Sex × Condition	0.41 (2.87)	12.20 (5.35)*
Sex × Block × Condition[a]	0.41 (1.49)	-12.22 (4.57)*
Random effects		
Subject		
Study 2		
Fixed effects	Coefficient (s.e.)	Coefficient (s.e.)
Intercept	-12.96 (2.40)**	-12.06 (2.28)**
Block	3.62 (1.23)**	2.46 (0.97)*
Condition	-6.21 (4.89)	-8.05 (4.97)
Block × Condition	5.76 (4.32)	7.69 (4.36)
Sex	-8.43 (3.67)*	-3.43 (3.36)
Sex × Block	7.56 (2.97)*	2.21 (2.24)
Sex × Condition	5.76 (6.32)	3.99 (6.73)
Sex × Block × Condition	-5.27 (5.59)	-3.53 (5.85)
Random effects		
Subject	321.2 (17.92)	429.7 (20.73)

Note. Condition coded 0 = no-feedback condition, 1 = feedback condition. Block coded 0 = pretest, 1 = posttest. Sex coded 0 = male 1 = female.

* $p < .05$, ** $p < .001$. [a] We broke down the final response's significant interaction; results indicated that the feedback was effective for males ($p < .001$) not for females ($p = .142$).

Table S5: Frequency of different types of justifications for the final bat-and-ball problem in Study 2

Justification	Feedback condition		No-feedback condition	
	Correct (n = 13)	Incorrect (n = 24)	Correct (n = 14)	Incorrect (n = 25)
Math - correct	11	-	14	-
Math - incorrect	-	16	-	18
Math - unspecified	-	1	-	2
Guess	1	2	-	-
Intuition	1	4	-	4
Other - correct	-	-	-	1
Other - incorrect	-	1	-	-
Other - unspecified	-	-	-	-

Note. Justification data of 4 participants is missing because their trial was excluded due to a missed deadline (see Exclusion Criteria).

Table S6: Average proportion of correct responses on the two transfer problems for improved versus unimproved reasoners

	n	*M*	*SE*
Feedback condition			
Improved reasoners	9	.56	.13
Unimproved reasoners	31	.44	.08
no ceiled pretest	27	.38	.08
ceiled pretest	4	.88	.13
No-feedback condition			
Improved reasoners	7	.64	.14
Unimproved reasoners	33	.50	.08
no ceiled pretest	25	.36	.08
ceiled pretest	8	.94	.07

Note. For the unimproved reasoners, we additionally distinguished between those who had no ceiled pretest performance (i.e., could still improve but simply did not) and those who had a ceiled pretest performance (i.e., were unable to improve).