The Smart System 1: evidence for the intuitive nature of correct responding on the bat-and-ball problem

Bence Bago, Wim de Neys

The smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem

Bence Bago
IAST – Toulouse School of Economics

Wim De Neys (wim.de-neys@parisdescartes.fr)
Paris Descartes University – Sorbonne Paris Cité
LaPsyDE (CNRS unit 8240)

Abstract

Influential work on reasoning and decision making has popularized the idea that sound reasoning requires correction of fast, intuitive thought processes by slower and more demanding deliberation. We present seven studies that question this corrective view of human thinking. We focused on the very problem that has been widely featured as the paradigmatic illustration of the corrective view, the well-known bat-and-ball problem. A two-response paradigm in which people were required to give an initial response under time-pressure and cognitive load allowed us to identify the presumed intuitive response that preceded the final response given after deliberation. Across our studies we observe that correct final responses are often non-corrective in nature. Many reasoners who manage to answer the bat-and-ball problem correctly after deliberation already solved it correctly when they reasoned under conditions that minimized deliberation in the initial response phase. This suggests that sound bat-and-ball reasoners do not necessarily need to deliberate to correct their intuitions, their intuitions are often already correct. Pace the corrective view, findings suggest that in these cases they deliberate to verify correct intuitive insights.

Keywords: Dual Process Theory; Heuristic Bias; Intuition; Deliberation; Two-Response Paradigm

Introduction

> "The intellect has little to do on the road to discovery. There comes a leap in consciousness, call it intuition or what you will, and the solution comes to you and you don't know why or how."
> - Albert Einstein (as quoted by Oesper, 1975)

> "It is through logic that we prove, but through intuition that we discover."
> - Henri Poincaré (as quoted by Poincaré, 1914)

There are few problems in the reasoning and decision making field that attracted so much interest as the bat-and-ball problem. In its original formulation as it was first proposed by Frederick (2005) the problem states:

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

Intuitively, the answer "10 cents" readily springs to mind. Indeed, this is the answer that most people tend to give (Bourgeois-Gironde & Van Der Henst, 2009; Frederick, 2005). However, although the answer seems obvious it is also dead wrong. Clearly, if the ball costs 10 cents and the bat costs $1 more, then the bat would cost $1.10. In this case, the bat and ball together would cost $1.20 and not $1.10. After some reflection it is clear that the ball must cost 5 cents and the bat costs – at a dollar more - $1.05 which gives us a total of $1.10.

Many people who are presented with the bat-and-ball problem will attest that the "10 cents" answer seems to pop-up in a split second whereas working to the "5 cents" solution seems to take more time and effort. As such, it is not surprising that the problem has been widely featured – from the scientific literature (Frederick, 2005; Kahneman, 2011; Mastrogiorgio & Petracca, 2014; Sinayev & Peters, 2015) to popular science best-sellers (Gladwell, 2005; Levitt & Dubner, 2010) to the pages of the Wall Street Journal (Lehrer, 2011) – as a textbook illustration of the dual process nature of human thinking.

The dual process model conceives of thinking as an interaction between intuitive and

(In Press) Thinking & Reasoning - doi: 10.1080/13546783.2018.1507949

deliberative processing, often referred to as System 1 and System 2 (e.g., Epstein, 1994; Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996). Although there are many types of dual process models they often postulate that the intuitive System 1 operates quickly and effortlessly whereas the deliberate System 2 is slower and effortful (i.e., it heavily burdens our limited cognitive resources)[1]. In the bat-and-ball problem it is System 1 that is assumed to be cueing the immediate "10 cents" response. Computation of the "5 cents" response is assumed to require the engagement of System 2 (Kahneman, 2011; Kahneman & Frederick, 2005). Because human reasoners have a strong tendency to minimize demanding computations, they will often refrain from engaging or completing the slow System 2 processing when the fast System 1 has already cued a response (Kahneman, 2011). Consequently, most reasoners will simply stick to the System 1 response that quickly came to mind and fail to consider the correct response. Reasoners who do manage to solve the problem correctly will need to correct the initially generated intuitive response after having completed their System 2 computations (Kahneman, 2011; Kahneman & Frederick, 2005; Morewedge & Kahneman, 2010).

At the heart of the dual process model of the bat-and-ball problem lies a "corrective" view on sound reasoning and deliberation: Correct responding is assumed to require correction of an intuitive System 1 response by slower and more demanding System 2 processing (Kahneman, 2011; Kahneman & Frederick, 2005). This can easily lead to a somewhat grim characterization of System 1 as a source of error that needs supervision from the deliberate System 2 (Marewski & Hoffrage, 2015). Bluntly put, System 2 is portrayed as the good guy that cleans up the mess left behind by the fast but error prone System 1. To be clear, it should be stressed that the dual process model does not simply postulate that intuitions are always incorrect. It is not disputed that intuitive responses can be appropriate and helpful in some cases (Evans, 2010; Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996). Likewise, it is also not claimed that deliberation will necessarily lead to a correct answer (Evans & Stanovich, 2013; Kahneman, 2011). The point is simply that intuitive responses have the potential to be incorrect and therefore need to be monitored and sometimes corrected. It is this corrective aspect of the model that can lead to the rather negative view of System 1. And it is this corrective nature of System 2 processing for which the bat-and-ball problem seems to present a paradigmatic example.

Although the corrective dual process model and related "intuition-as-bias" view have been very influential various scholars have called for a more positive view towards intuitive processing (e.g., Klein, 2004; Peters, 2012; Reyna, 2012). More generally, as our opening quotes illustrate, historically speaking, leading mathematicians and physicist have long favored a quite different look on intuition and deliberation. In scientific discoveries and mathematical breakthroughs intuition has often been conceived as guiding the intellect. Great minds such as Newton, Einstein, Poincaré, and Kukulé famously reported how their major breakthroughs – from the discovery of gravity to the Benzene ring structure – came to them intuitively (Ellenberg, 2015; Marewski & Hoffrage, 2015). Although this "intuition-as-a-guide" view does not deny that more effortful, deliberate thinking is indispensable to validate and develop an initial intuitive idea, the critical origin of the insight is believed to rely on intuitive processes (Ellenberg, 2015; Marewski & Hoffrage, 2015). In contrast with the corrective dual process view, here sound reasoning is not conceived as a process that corrects erroneous intuitions but as a process that builds on correct insights.

The present study focuses on an empirical test of the corrective nature of sound reasoning in the bat-and-ball problem. Our

[1] Operation speed and effort are typical correlates of System 1 and 2 processing. The idea is that these features have often been associated with System 1 and 2 processing. But this does not necessarily need to be the case; the features do not necessarily need to align (e.g., a process might be effortless but slow, e.g., "incubation", Gilhooley, 2016), and other features can be proposed to differentiate System 1 and 2 processing (e.g., "autonomy", Pennycook, 2017). See Evans and Stanovich (2013), and the debate between Melnikoff and Bargh (2018) and Pennycook, De Neys, Evans, Stanovich, and Thompson (2018), for an extensive discussion.

motivation was that although the problem is considered a paradigmatic example of the need to correct an initial intuitive response and has been widely used to promote this view, the critical corrective assumption has received surprisingly little direct empirical testing. Arguably, one of the reasons is that the characterization seems self-evident from an introspective point of view. As we already noted, few people (including the present authors) would contest that it feels as if the "10 cents" answer pops up immediately and arriving at the "5 cents" solution requires more time and effort. Indeed, many scholars have referred to this introspective experience to warrant the corrective assumption (Frederick, 2005; Kahneman, 2011). However, while introspection is not without its merits, it is also well established that introspective impressions can be deceptive (Mega & Volz, 2014; Schooler, Ohlsson, & Brooks, 1993).

In addition to introspection, Frederick (2005) also cited some indirect support for the corrective assumption in the paper that introduced the bat-and-ball problem. For example, Frederick observed that incorrect responders rate the problem as easier than correct responders and suggests that this presumably indicates that correct responders are more likely to consider both responses (see also Mata & Almeida, 2014; but see also Szaszi, Szollosi, Palfi, & Aczel, 2017). The problem is that although such assumptions are not unreasonable, they do not present conclusive evidence. Clearly, even when the assumption holds that correct responders are more likely to consider both the incorrect and correct responses, it does obviously not imply that they considered the incorrect response *before* the correct response.

Other potential support comes from latency studies. A number of studies reported that correct "5 cents" responses take considerably longer than incorrect "10 cents" responses (e.g., Alós-Ferrer, Garagnani, & Hügelschäfer, 2016; Johnson, Tubau, & De Neys, 2016; Stupple, Pitchford, Ball, Hunt, & Steel, 2017; Travers, Rolison, & Feeney, 2016). For example, in one of our own studies we observed that correct responders needed on average about a minute and a half to enter their response whereas incorrect responders only took about 30 seconds (Johnson et al., 2016). Although this fits with the claim that System 2 processing is slower than System 1 processing, it does not necessarily imply that someone who engaged in System 2 reasoning, first engaged in System 1. That is, the fact that a correct response takes more time does not imply that correct responders generated the incorrect response before they considered the correct response. In theory, they might have needed more time to complete the System 2 computations without ever having considered the incorrect response. Hence, if we want to obtain solid evidence for the corrective assumption we need to move beyond mere latency data.

Somewhat more convincing evidence for the corrective dual process assumption in the bat-and-ball problem comes from a recent paper by Travers et al. (2016). In the study Travers et al. adopted a mouse tracking paradigm. In this paradigm different response options are presented in each of the corners of the screen (e.g., "10 cents", "5 cents") and participants have to move the mouse pointer from the center of the screen towards the response option of their choice to indicate their decision. This procedure can be used to study the time-course of decisions on the basis of participant's mouse cursor trajectories (e.g., Spivey, Grosjean, & Knoblich, 2005). For example, do reasoners who ultimately select the correct response tend to move the mouse first towards the incorrect response? Travers et al. found that this was indeed the case. After about 5s participants started to make initial movements towards the incorrect "10 cents" option. However, movements towards the correct response were not observed until about 5 s later. These findings present some support but they are not conclusive. Note that if a response is truly intuitive, one might expect it to be cued instantly upon reading the problem. In this sense, the 5 s time lag before participants started to make mouse movements in the Travers et al. study is still quite long. This leaves open the possibility that the procedure is not picking up on earlier intuitive processing (Travers et al., 2016).

In the present paper we report a research project involving a total of seven studies in which we aimed to test the corrective nature of correct responding in the bat-and-ball problem directly. In other words, we aim to test the role of effortful thinking in generating the correct "5 cents" response. We therefore adopted the two-response

paradigm (Thompson, Prowse Turner, & Pennycook, 2011). Thompson and colleagues developed this procedure to gain direct behavioral insight into the timing of intuitive and deliberative response generation. In the paradigm participants are presented with a reasoning problem and are instructed to respond as quickly as possible with the first, intuitive response that comes to mind. Subsequently, they are presented with the problem again, and they are given as much time as they want to think about it and give a final answer. Interestingly, a key observation for our present purposes was that Thompson and colleagues observed that people rarely change their initial response in the deliberation phase (Pennycook & Thompson, 2012; Thompson & Johnson, 2014; Thompson et al., 2011). This lack of answer change tentatively suggests that in those cases where a correct response was given as final response, the very same response was generated from the start. In other words, the correct response might have been generated fast and intuitively based on mere System 1 processing (Pennycook & Thompson, 2012; Thompson & Johnson, 2014; see also Bago & De Neys, 2017, and Newman, Gibb, & Thompson, 2017).

However, past two-response studies used problems which were typically considerably easier than the bat-and-ball problem (Travers et al., 2016). Note that the dual process model does not entail that correct responding requires System 2 thinking in all possible situations and conditions. In some elementary tasks (e.g., Modus Ponens inferences in conditional reasoning, see Evans, 2010) the processing needed to arrive at the correct response might be so basic that it has been fully automatized and incorporated as a System 1 response. Likewise, in some cases System 1 might not generate a (incorrect) response and there will obviously be no need to correct it. Hence, findings pointing to correct intuitive responding might be attributed to the exceptional, non-representative nature of the task (Aczel, Szollosi, & Bago, 2016; Mata, Ferreira, Voss, & Kollei, 2017; Pennycook, Fugelsang, & Koehler, 2012; Singmann, Klauer, & Kellen, 2014; Travers et al., 2016). Proponents of the corrective dual process view can still argue that in prototypical cases – with the bat-and-ball problem as paradigmatic example - correct responding can only occur after deliberation and correction of an intuitive response. In addition, one might argue that at least in the initial two-response studies participants were simply instructed - and not forced - to respond intuitively. Hence, participants might have failed to respect the instructions and ended up with a correct first response precisely because they engaged in System 2 processing. Clearly, one has to try to make maximally sure that only System 1 is engaged at the initial response phase.

In the present study we adopt the two-response paradigm in a reasoning task with items that are directly modeled after the bat-and-ball problem. We also use stringent procedures to guarantee that the first response is intuitive in nature. Participants are forced to give their first response within a challenging deadline (e.g., 4 s in Study 1, the time needed to simply read the problem as indicated by pretesting). In addition, during the initial response phase participants' cognitive resources are also burdened with a secondary load task. The rationale is simple. System 2 processing, in contrast with System 1, is often conceived as time and resource demanding (Kahneman, 2011; Kahneman & Frederick, 2005). By depriving participants from these resources we attempt to "knock" out System 2 as much as possible during the initial response phase (Bago & De Neys, 2017). In other words, by having participants reason under conditions that minimize possible deliberation in the initial response phase, we attempt to identify the presumed intuitive response that precedes the final response given after deliberation. Finally, we also use a range of procedures to eliminate possible confounds resulting from task familiarity or guessing.

To give away the punchline, our key result is that although we replicate the biased responding on the bat-and-ball problem, we also find consistent evidence for correct "intuitive" responding in the initial response phase. Whenever people manage to give the correct "5 cent" answer as their final response after deliberation, they often already selected this answer as their initial response when possible deliberation was experimentally minimized. In the different studies we use various manipulations (response format variations, Study 1-5; response justification elicitation, Study 6-7)

that help to pinpoint the nature of the intuitive (i.e., initial) correct responses and the contrast with deliberate correct responses after reflection. Based on our empirical findings we will argue that the role of System 1 and 2 in dual process theories might need to be re-conceptualized: In line with the "intuition-as-a-guide" view favored by Einstein and Poincaré, it seems that in addition to the correction of an incorrect intuition, deliberation is often also used to verify and justify a correct intuitive insight.

Study 1

Method

Participants

In Study 1, 101 Hungarian undergraduate students (87 female, Mean age = 19.8 years, SD = 1.5 years) from the Eotvos Lorand University of Budapest were tested. Only freshmen were allowed to participate, so their highest completed educational level was high school (except 1 subject who reported that she already had a Bachelor degree). Participants received course credit for taking part. Participants in Study 1 (and all other reported studies) completed the study online.

Materials

Reasoning problems. In total, eight content modified version of the bat-and-ball problem were presented. The original bat-and-ball problem is frequently featured in scientific studies and popular science writing which implies that prospective participants might be familiar with it (Haigh, 2016). Previous studies have shown that more experienced participants (who took part in more studies and thus had a higher likelihood to have previously seen or solved the bat-and-ball problem) performed significantly better than naïve participants (Stieger & Reips, 2016). Furthermore, Chandler, Mueller, and Paolacci (2014) have also found a positive correlation between the performance on the Cognitive Reflection Test (CRT, a short 3-item questionnaire that includes the bat-and-ball problem) and the number of experiments people participated in. However, this correlation disappeared when the researchers used structurally identical, but content modified version of the problems. We used similar content modified versions of the original bat-and-ball problem (e.g., problems stated that a cheese and a bread together cost 2.90 euro[2] or that an apple and an orange together cost 1.80 euro) in the present study to minimize the effect of familiarity or prior exposure on task performance.

Furthermore, we presented multiple items to help us test for a possible guessing confound. Study 1 adopted a binary response format (see further). Hence, mere guessing at the initial response stage would already result in 50% correct "intuitive" responses. However, by presenting multiple items and computing a measure of "stability" (i.e., out of the presented items, on how many did the participant respond similarly?) we can control for the guessing account. If System 1 computes the correct response one would expect that reasoners manage to select it consistently. If correct responding results from mere guessing, it should only by observed on about half of the presented items. Note that in theory presenting multiple items might also boost participants' accuracy because of a learning or transfer effect: after having solved the problem once, participants might have learned to apply the same strategy on subsequent items. However, Chandler et al.'s (2016) work suggests that people's knowledge about the problem is highly item and content specific. Hoover and Healy (2017) also failed to observe transfer effects with repeated presentation of (content modified) problems. Hence, by changing the content of every item we can expect to minimize learning effects.

In Study 1 each problem was always presented with two answer options; the correct response (e.g., "5 cents" in the original bat-and-ball problem) which is assumed to require System 2 deliberation and the "heuristic" response which is assumed to be cued by System 1 (e.g., "10 cents" in the original problem). We will use the labels correct and "heuristic" response to refer to these answers options. Mathematically speaking, the correct equation to solve the bat-and-ball problem is: 100 + 2x =

[2] Note that in all our studies with Hungarian subjects we used euro instead of dollar units since this currency is more familiar to them.

110, instead, people are thought to be intuitively using the "100 + x = 110" 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 (e.g., if the two objects were said to cost 2.30 in total, the presented heuristic response option was 30 cents and the presented correct response option was 15 cents).

Participants had to indicate their answer by clicking on one of the options with the mouse. After providing an answer, they immediately advanced to the next problem. In order to minimize the influence of reading times and get a less noisy measure of reasoning time, problems were presented serially. First, the first sentence of the problem was presented, which always stated the two objects and their cost together (e.g., "An apple and an orange cost 1.80 euros in total"). Next, the rest of the problem was presented under the first sentence (which stayed on the screen), with the question and the possible answer options. The following illustrates the full problem format:

> An apple and an orange cost 1.80 euros in total.
> The apple costs 1 euro more than the orange.
> How much does the orange cost?
> ○ 40 cents
> ○ 80 cents

To further assure that possible correct (or incorrect) responses did not originate from guessing we also presented control problems (see De Neys, Rossi, & Houdé, 2013; Travers et al., 2016). In the regular bat-and-ball problem the intuitively cued response is assumed to cue an answer that conflicts with the correct answer. In the "no-conflict" control problems, the conflict was removed. This was achieved by deleting the critical relational "more than" statement. With the above example, a control problem looked as follows:

> An apple and an orange cost 1.80 euros in total.
> The apple costs 1 euro.
> How much does the orange cost?
> ○ 40 cents
> ○ 80 cents

In this case the intuitively cued "80 cents" answer was also correct. The second response option for the control problems was always the correct response divided by 2 (e.g., "40 cents" in the example). Half of the presented problems were regular "conflict" problems, half of them were control problems. If participants are not guessing, performance on the control problems should be at ceiling (e.g., De Neys et al., 2013).

Two problem sets were used in order to counterbalance the item content; the conflict items in one set were the control items in the other, and vice-versa. Participants were randomly assigned to one of the sets. This counterbalancing minimized the possibility that item contents would affect the contrast between conflict and control items. The presentation order of the items was randomized in both sets. All problems are presented in the Supplementary Material, section A.

Load task. We wanted to try to make maximally sure that participants' initial response was intuitive (i.e., System 2 engagement is minimized). Therefore, we used a cognitive load task (i.e., the dot memorization task, see Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001) to burden participants' cognitive resources The rationale behind the load manipulation is simple; System 2 processing is often assumed to require executive cognitive resources, while it is assumed that System 1 processing does not require these resources (Evans & Stanovich, 2013). Consequently, if we burden someone's executive resources while they are asked to solve reasoning problems, System 2 engagement is less likely. We opted for the dot memorization task because it is well-established that it successfully burdens participant's executive resources (De Neys & Schaeken, 2007; De Neys & Verschueren, 2006; Franssens & De Neys, 2009; Johnson, et al., 2016; Miyake et al., 2001). Before each reasoning problem participants were presented with a 3 x 3 grid, in which 4 dots were placed. Participants were instructed that it was critical to memorize it even though it might be hard while solving the reasoning problem. After answering the reasoning problem participants were shown four different matrixes and they had to choose the correct, to-be-memorized pattern. They received

feedback as to whether they chose the correct or incorrect pattern. The load was only applied during the initial response stage and not during the subsequent final response stage in which participants were allowed to deliberate (see further).

Procedure

Reading pre-test. Before we ran the main study we also recruited an independent sample of 64 participants for a pre-test in which participants were simply asked to read our reasoning task items and randomly click on a response option when they were ready. The idea was to base the response deadline in the main reasoning task on the average reading time in the reading test. Note that dual process theories are highly underspecified in many aspects (Kruglanski, 2013); they argue that System 1 is faster than System 2, but do not further specify how fast System 1 is exactly (e.g., System 1 < x seconds). Hence, the theory gives us no unequivocal criterion on which we can base our deadline. Our "average reading time" criterion provides a practical solution to define the response deadline. If people are allotted the time they need to simply read the problem, we can assume that System 2 engagement is minimal. Full procedural details are presented in the Supplementary Material, Section D. The average reading time of the sample was 3.87 s (SD = 2.18 s). To give participants some minimal leeway, we rounded the average reading time to the closest higher natural number and set the response deadline to 4 seconds.

Reasoning task. The experiment was run online. Participants were specifically instructed at the beginning that we were interested in their very first, initial answer that came to mind. They were also told that they would have additional time afterwards to reflect on the problem and could take as much time as they needed to provide a final answer. The literal instructions that were used, stated the following (translated from Hungarian):

> **Welcome to the experiment!**
> **Please read these instructions carefully!**
>
> This experiment is composed of 8 questions and a couple of practice questions. It will take about 10 minutes to complete and it demands your full attention. You can only do this experiment once.
>
> In this task we'll present you with a set of reasoning problems. We want to know what your initial, intuitive response to these problems is and how you respond after you have thought about the problem for some more time. Hence, as soon as the problem is presented, we will ask you to enter your initial response. We want you to respond with the very first answer that comes to mind. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible. Next, the problem will be presented again and you can take all the time you want to actively reflect on it. Once you have made up your mind you enter your final response. You will have as much time as you need to indicate your second response.
>
> After you have entered your first and final answer we will also ask you to indicate your confidence in the correctness of your response.
>
> In sum, keep in mind that it is really crucial that you give your first, initial response as fast as possible. Afterwards, you can take as much time as you want to reflect on the problem and select your final response.
>
> You will receive 500 HUF for completing this experiment.
>
> Please confirm below that you read these instructions carefully and then press the "Next" button.

After this general introduction, participants were presented with a task specific introduction which explained them the upcoming task and informed them about the response deadline. The literal instructions were as follows:

> We are going to start with a couple of practice problems. First, a fixation cross will appear. Then, the first sentence of the problem is going to be presented for 2 seconds. Next, the rest of the problem will be presented.

> As we told you we are interested in your initial, intuitive response. First, we want you to respond with the very first answer that comes to mind. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible. To assure this, a time limit was set for the first response, which is going to be 4 seconds. When there is 1 second left, the background colour will turn to yellow to let you know that the deadline is approaching. Please make sure to answer before the deadline passes. Next, the problem will be presented again and you can take all the time you want to actively reflect on it. Once you have made up your mind you enter your final response.
>
> After you made your choice and clicked on it, you will be automatically taken to the next page.
> After you have entered your first and final answer we will also ask you to indicate your confidence in the correctness of your response.
> Press "Next" if you are ready to start the practice session!

After the specific instruction page participants solved two unrelated practice reasoning problems to familiarize them with the procedure. Next, they solved two practice dot matrix practice problems (without concurrent reasoning problem). Finally, at the end of the practice, they had to solve the two earlier practice reasoning problems under cognitive load.

Each problem started with the presentation of a fixation cross for 1000 ms. After the fixation cross disappeared, the dot matrix appeared and stayed on the screen for 2000 ms. Then the first sentence appeared alone for 2000 ms. Finally, the remaining part of the problem appeared (while the first sentence stayed on screen). At this point participants had 4000 ms to give an answer; after 3000 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. The position of the correct answer alternative (i.e., first or second response option) was randomly determined for each item.

After the initial response, participants were asked to enter their confidence in the correctness of their answer on a scale from 0% to 100%, with the following question: “How confident are you in your answer? Please type a number from 0 (absolutely not confident) to 100 (absolutely confident)”.

After indicating their confidence, they were presented with four dot matrix options, 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 dot pattern on subsequent trials.

Finally, the same item was presented again, and participants were asked to provide a final response. Once they clicked on one of the answer options they were automatically advanced to the next page where they had to provide their confidence level again.

The colour of the answer options was green during the first response, and blue during the final response phase, to visually remind participants 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 coloured 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.

Exclusion criteria. In total, 26.7% (n = 27) of participants reported they had seen the original bat-and-ball problem before. Thirteen participants in this group managed to give the correct “5 cents” response. Although we used content modified problem versions in our study, we wanted to completely eliminate the possibility that these participants’ prior knowledge of the original correct solution would affect the findings. Therefore, we decided to discard all data from the participants who had seen the original bat-and-ball problem before and knew

the correct solution (i.e., 13% from total sample, 88 participants were further analyzed).

The remaining participants failed to provide a first response before the deadline in 10.3% of the trials. In addition, in 11.3% of the trials participants responded incorrectly to the dot memorization load task. All these trials were removed from the analysis because it cannot be guaranteed that the initial response resulted from mere System 1 processing: If participants took longer than the deadline, they might have engaged in deliberation. If they failed the load task, we cannot be sure that they tried to memorize the dot pattern and System 2 was successfully burdened. In these cases we cannot claim that possible correct responding at the initial response stage is intuitive in nature. Hence, removing trials that did not meet the inclusion criteria gives us the purest possible test of our hypothesis.

In total, 18.3% of trials were excluded and 575 trials (out of 704) were further analyzed (initial and final response for the same item counted as 1 trial). For completeness, note that in Study 1 - and all other studies reported here - we also ran our analyses with all trials included. Key findings were never affected.

Statistical analyses. Throughout the article we used mixed-effect regression models (Baayen, Davidson, & Bates, 2008) in which participants and items were entered as random intercepts to analyse our results. For the binary choice data we used logistic regression while for the continuous confidence and reaction time data we used linear regression.

Results

Table 1 gives an overview of the results. For consistency with previous work we first focus on the response accuracies for the final response. In line with the literature we find that most participants fail to solve the conflict versions of the bat-and-ball problem correctly. On average, final accuracy on the conflict items reached 24.5% (SD = 43.1). As one might expect, on the no-conflict control problems where System 1 processing is hypothesized to cue the correct response, final accuracy is at ceiling with about 96% (SD = 19.6) correct responses. These final response accuracies are consistent with what can be expected based on previous work that adopted a classic one-response paradigm (e.g., 21.6% and 97.4% in Johnson et al., 2016; 21% and 98% in De Neys et al., 2013). This indicates that the use of a two-response paradigm does not distort reasoning performance per se (Thompson et al., 2011).

The initial response accuracies are more surprising. In about 20.1% (SD = 40.2%) of the conflict trials people already give the correct answer as their initial response. This suggests that participants can solve the bat-and-ball problem “intuitively". However, the raw percentage of correct intuitive responses is not fully informative. We can obtain a deeper insight into the results by performing a Direction of Change analysis on the conflict trials (Bago & De Neys, 2017). This means that we look at the way a given person in a specific trial changed (or didn’t change) her initial answer after the deliberation phase. More specifically, people can give a correct or incorrect response in each of the two response stages. Hence, in theory this can result in four different types of answer change patterns (“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). According to the standard dual process model, two types of direction of change should be most common; 00 and 01. The “00” category implies that System 1 processes generated an initial incorrect response and System 2 thinking did not manage to override it. Consequently, the person is biased. The “01” category presents the standard correction case; System 1 processing will generate an initial incorrect response, but System 2 processing later manages to correct it in the final response stage.

Table 2 summarizes the frequencies of each direction of change category for the conflict problems. The most prevalent category is the 00 one; this pattern was generated in 71.8% of the trials. This means that both at the initial and final response stage participants were typically biased when solving our problems, which mirrors the overall accuracy pattern. One can also observe that there is a non-negligible amount of 01 responses (8.1% of trials). This is in accordance with the dual process predictions. Some reasoners

initially generated the incorrect response, but managed to correct it after deliberation. However, the problem is that we observe about twice as many 11 cases (16.5% of trials). This means that in many cases (i.e., 67.2%) in which reasoners managed to give the correct response as their final answer, they already gave it "intuitively" at the initial response stage. We refer to this critical number [(i.e., 11/(11+01) ratio] as the % of non-corrective correct responses or non-correction rate in short. Overall, this result implies that contrary to the core dual process assumption, correct responding in the bat-and-ball problem does not necessarily require System 2 correction. System 2 correction does exist (i.e., we observe some 01 cases) but the point is that this correction seems to be far less common than assumed.

For completeness, Table S3 in the Supplementary Material also gives an overview of the direction of change findings for the no-conflict control items. Not surprisingly, here responses predominantly (93%) fell in the 11 category.

Clearly, a critic might argue that given our binary response format our correct initial responses on the conflict items result from mere guessing. Indeed, our task is quite challenging – people have to respond within a very strict deadline and under cognitive load. In theory, it is possible that participants found the task too hard and just randomly clicked on one of the presented solutions. However, the ceiled initial performance on the no-conflict control problems argues against a general guessing confound (i.e., 93.4% vs 20.1% correct initial responses on the no-conflict and conflict problems respectively, χ^2 (1) = 56.64, $p < 0.0001$, b = -4.77). If our task was so challenging that participants had to guess because they were not even able to read the problem information, their performance on the conflict and no-conflict problems should not have differed and should have hovered around 50% in both cases. Further evidence against a guessing account is also provided by our stability analysis (see below).

Our direction of change analysis was computed across items and participants. One might wonder whether participants are stable in their preference for one or the other type of change category. That is, does an individual who produces a correct (incorrect) initial response on one conflict problem does so consistently for the other items? To answer this question we calculated for every participant on how many conflict problems they displayed the same direction of change category. We refer to this measure as the *stability index*. For example, if an individual shows the same type of direction of change on all four conflict problems, the stability index would be 100%. If the same direction of change is only observed on two trials, the stability index would be 50% etc. Table 3 presents an overview of the findings. Note that due to our methodological restrictions (discarding of answers after the deadline and for which the load memorization was not successful) for a small number of participants less than four responses were available. Here the stability index is calculated over the available items.

As Table 3 shows, the dominant category is the 100% stability one. The average stability index on conflict items was 87.1% (SD = 20.5)[3]. This indicates that the type of change is highly stable at the individual level. If people show a specific direction of change pattern on one conflict problem, they tend to show it on all conflict problems. Note that the stability index directly argues against the guessing account. If people were guessing when giving their initial answer, they should not tend to pick the same response consistently. Likewise, the high stability also argues against a possible practice or learning account. Although we used content-modified items (i.e., each item mentioned a unique set of products and total dollar amount), one might argue that the repeated presentation of multiple problems helped participants to learn and automatize the calculation of the correct response. Correct intuitive responding would arise near the end of the study, but would not be observed at the start. However, pace the learning account, the high stability indicates that participants' performance did not change across the study. If participants managed to give a correct initial answer at the end of the study, they already did so at the start. If they were biased at

[3] Table S4 in the Supplementary Material shows that the stability was also high on the control problems with an average of 93% (SD = 12.8).

the start, they were so at the end. This directly argues against a learning account.

Discussion

Consistent with the prior literature, Study 1 showed that people are typically biased and fail to solve the bat-and-ball problem. Considered in isolation, this fits with the standard dual process story that solving the problem is hard and requires deliberate System 2 processing to override an incorrect intuition. However, the key finding is that in those cases where people do manage to give a correct final response after deliberation, they often already selected this answer as their first, intuitive response. Even though we experimentally reduced the use of System 2 deliberation and forced people to respond in a mere 4 s, correct responders often also gave the correct answer as their initial response. In many cases correct responses were non-corrective in nature. In other words, this suggests that correct responders do not necessarily need to correct their intuition, their intuition is often already correct.

The high non-correction rate directly argues against the dual process theory assumption concerning the corrective nature of System 2 processing. But obviously, the results are based on one single study. In Study 2-5 we therefore tested the robustness of the findings. We examined whether we can replicate the pattern and whether it is robust to variations in response format. In Study 1 we used a binary forced-choice response format. In Study 2 and 3 we also introduced a condition with a 4-option response format. In Study 4 and 5 we used a free-response format. Each response format has some methodological advantages and disadvantages. The simple binary format allows us to set the most stringent deadline for the initial response stage. If people are presented with and have to read more response options or have to type their own response, they will need more time to enter a response. On the other hand, the multiple response and free response format allow better control against guessing and provide insight into the nature or specificity of the intuitive response. Clearly, if an individual selects a correct initial response in a binary choice format, this does not necessarily imply that she has calculated that the correct response is "5 cents". She might have intuitively detected that the "10 cents" cannot be right (e.g., because the sum would be larger than $1.10) without knowing the correct solution. If participants select the "5 cents" response from a wider range of options – or generate it themselves – we can conclude that they also computed the correct response. In other words, we can test how precise their intuitive knowledge is.

In Study 1 we tested Hungarian undergraduates. In Study 2-5, we also recruited participants from a wider range of populations (e.g., in terms of nationality, age, and education level).

Study 2-5

For ease of presentation we will present a single results section in which the response format factor is included as factor in the analyses. Here we present an overview of the method sections of Study 2-5.

Method – Study 2: 2-option vs 4-option format (crowdsource sample)

Participants

A total of 372 participants were tested (196 female, Mean age = 39.6 years, SD = 13.5 years). Participants were recruited on-line via the Crowdflower platform and received $0.20 for their participation. Only native English speakers from the USA or Canada were allowed to take part in the experiment. A total of 36% of the participants reported high school as highest completed educational level, while 62% reported having a post-secondary education degree (2% reported less than high school, and 1 participant did not provide this information).

Materials and Procedure

Reading pre-test. Half of the participants were presented with four response options in Study 2. Since reading through more options will in itself take more time, we decided to run a new reading pre-test with the 4-option format (see supplementary material, section D, for full details). The mean reading time in the pre-test sample was 4.3 s (SD = 2 s). As in Study 1, we rounded the deadline to the nearest higher natural

number. Hence, the time limit in the 4-option format was set to 5 s (vs 4 s in the 2-option format).

Reasoning task. Participants were randomly allocated to the 2-option or 4-option treatment. The 2-option condition was completely identical to Study 1 except for the fact that material was presented in English and not in Hungarian. In the 4-options condition, two foil response options were presented in addition to the heuristic and correct response option. We used a "high" and a "low" foil option and used the following rules to determine them: the "high" foil was always the sum of the heuristic and correct options, whereas the "low" foil was always the greatest common divisor of the correct and heuristic option that was smaller than the correct answer. For example, in the original bat-and-ball problem, these would be the four response options: 1 (low foil), 5 (correct), 10 (heuristic), 15 (high foil).

The presentation order of the response options was always the same in the initial and final response stages, but was counterbalanced across trials. All problems with their respective answer options are presented in the Supplementary Material, section A.

Exclusion criteria. The same exclusion criteria were applied as in Study 1. In total 15.3% of participants were excluded because they had seen the bat-and-ball problem before and knew the correct response. We further excluded all trials where participants failed to provide a response within the deadline (7.1% of trials; 7% in the 2-option and 7.3% in the 4-option condition) or did not provide the correct response to the load memorization task (13.3% of trials; 14.9% in the 2-option and 11.5% in the 4-option condition). Altogether, 18.7% of trials were excluded and 2049 trials (out of 2520) were further analyzed.

Method - Study 3: 2-option vs 4-option format (Hungarian student sample)

Participants

In total, 121 Hungarian university students from the Eotvos Lorand University of Budapest were tested (92 female, Mean age = 22.2 years, SD = 1.4 years). Participants received 500 HUF (~$1.7) for taking part. In total, 83% of subjects reported high school as highest educational level, and 17% reported that they already obtained a post-secondary educational degree.

Materials and procedure

We used Hungarian translations of the material but otherwise the Study 3 design was identical to Study 2. The same exclusion criteria were applied. In total 29.8% participants were excluded (85 were further analysed) because they had seen the original bat-and ball problem before and knew the correct answer. We further excluded trials where participants failed to provide a response within the deadline (9.4% of trials; 8.3% in the 2-option, and 10.5% in the 4-option condition) or did not provide the correct response to the memorization load task (11.9% of trials; 11.9% in both the 2-option and 4-option condition). Altogether, 19.6% of trials were excluded and 547 trials (out of 680) were further analyzed.

Method – Study 4: free response format (crowdsource sample)

Participants

A total of 47 participants took part (30 female, Mean age = 43.8 years, SD = 15.2 years) in this study. Participants were recruited via Crowdflower and received $0.20 for completing the study. 32% of participants reported high school as highest completed educational level, and 66% reported having a post-secondary educational level degree (2% reported less then high school).

Materials and procedure

Study 4 used a free response format. 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. Obviously, this procedure will take longer than simple response selection and will be affected by variance in typing skill. To avoid unwarranted response rejection we decided to set a liberal deadline of 10 s in the initial response stage. The problem background turned yellow 2

seconds before the deadline. Note that previous studies that adopted a free response format without time-restrictions reported average response latencies for correct answers of over 30 s (Stupple et al., 2017; Johnson et al., 2016). Hence, by all means the 10 s deadline remains challenging. In addition, participants still had to give their initial response under secondary task load. Consequently, even with the longer deadline we still minimize System 2 engagement during the initial response stage. Otherwise, the study design was completely similar to Study 1.

The same exclusion criteria were applied as in Study 1. In total 14.9% participants were excluded because they had seen the original bat-and-ball problem before and knew the correct response. We further excluded trials where participants failed to provide a response within the deadline (2.5% of trials) or did not provide the correct response to the memorization load task (12.8% of trials). Altogether, 14.4% of trials were excluded and 274 trials (out of 320) were further analyzed.

Method – Study 5: free response format (Hungarian convenience sample)

Participants

A total of 55 Hungarian volunteers participated (50 female, Mean age = 33.8 years, SD = 9.3 years) in this study. Participants were recruited online through the help of social media. Participants completed the experiment online. 33% of participants reported that their highest completed educational degree was high school, while 64% reported having a post-secondary educational level degree (4% reported less then high school).

Materials and procedure

The same procedure was used as in Study 4 but with a stricter initial response deadline of 8 s. This was based on the fact that we observed that it took Study 4 participants on average only 5.1 s (SD = 1.46 s) to enter their initial response. The problem background again turned yellow 2 seconds before the deadline.

The same exclusion criteria were applied as in Study 1. In total 18.2% of participants were excluded because they had seen the original bat-and-ball problem before and knew the correct response. We further excluded trials where participants failed to provide a response within the deadline (11.1% of trials) or did not provide the correct response to the memorization load task (11.7% of trials). Altogether, 19.7% of trials were excluded and 289 trials (out of 360) were further analyzed.

Results

Table 1 and 2 give an overview of the accuracy and direction of change findings in each study. The bottom rows of the tables show the overall average and the average in function of the three response formats we used across Study 1-5. The overall average sketches a pattern that is consistent with the Study 1 findings. Most people are biased when solving the conflict bat-and-ball problems with initial and final accuracies of 13.8% and 16.8% whereas performance on the control no-conflict versions is at ceiling. Direction of change results for the conflict problems show that there is some correction going on with 5.2% of "01" cases in which an initial incorrect response is corrected after deliberation in the final response stage. However, the key finding is that the "11" cases in which a correct final response is preceded by a correct initial response are twice as likely (10.9%). Indeed, the overall non-correction rate reached 67.6% - which is virtually identical to the 67.2% rate observed in Study 1. When eyeballing the averages for the three response formats separately, it is clear that by and large this overall pattern is observed with each format. There is no evidence for a systematic decrease in initially correct responding in the 4-option and free response studies. Statistical analyses showed that the initial accuracy, χ^2 (2) = 1.11, p = 0.57, final accuracy, χ^2 (2) = 0.67, p = 0.72, as well as the rate of non-corrective correct responding, χ^2 (2) = 4.2, p = 0.12, were not significantly affected by the response format. This implies that our core finding is robust: Across multiple studies with different response formats (and a range of populations) we consistently observe that when reasoners solve the bat-and-ball problem correctly, they typically give the same answer as their first, intuitive response. This questions the assumption that System 2 is needed to correct our intuition.

As a side note, an intriguing observation is that when looking at the individual studies, one might note that in the two studies with participants recruited on the Crowdflower platform (Study 2: 2-option and 4-option formats and Study 4: free response format) we observed overall lower accuracies than in the other studies, both at the initial and final response stage. This trend reached significance at the initial, χ^2 (1) = 7.1, p = 0.008, b = 1.44, and at the final response stage, χ^2 (1) = 4.35, p = 0.04, b = 1.5, as well. Nevertheless, despite the overall lower accuracy we observe the same key pattern. The non-correction rate did not differ significantly in our Crowdflower studies and the other studies, χ^2 (1) = 1.63, p = 0.20.

Table 3 gives an overview of the stability index on the conflict items in Study 2-5. As in Study 1 we observe that the index is consistently high with an overall average value of 91.3%. This indicates that the direction of change pattern is highly stable at the individual level. If people show a specific direction of change pattern on one conflict problem, they show this same pattern on all conflict problems. This argues against a guessing and practice account. If participants gave correct intuitive responses because they guessed or because the repeated presentation allowed them to automatize the calculations after the first trial(s), their performance should have shown more variability[4]. Nevertheless, there was some variability and especially with respect to the practice account one might argue that it can be informative to focus exclusively on the very first problem that reasoners solved. In an additional analysis we therefore included only the first conflict problem that reasoners solved and excluded all later trials. Obviously, given that we restrict the analysis to a single trial with only a small number of critical correct (initial and final) responses per study, it should not be surprising that the data are noisier. Nevertheless, key finding is that even in this single trial analysis, the overall non-correction rate across our studies still reached 42% (individual experiments range from 11% to 75%; average 2-response format = 30.1%, average 4-response = 61.5%, average free response = 40%, see supplementary Table S7 for a full overview). Although the effect is less robust than in the full analysis, this confirms that the critical correct initial responding is present from the start.

Having established that the core finding concerning the generation of a correct System 1 intuition is robust, we can dig somewhat deeper into the findings. One interesting open question is whether correct initial responders are faced with two competing intuitions at the first response stage. That is, a possible reason for why people in the 11 category manage to give a correct initial response might be that the problem simply does not generate an intuitive heuristic "10 cents" response for them. Hence, they would only generate a correct "5 cents" intuition and would not be faced with an interfering heuristic one. Alternatively, they might generate two competing intuitions, but the correct intuition might be stronger and therefore dominate (Bago & De Neys, 2017).

We can address this question by looking at the contrast between conflict and no-conflict control problems. If conflict problems cue two conflicting initial intuitive responses, people should process the problems differently than the no-conflict problems (in which such conflict is absent) in the initial response stage. Studies on conflict detection during reasoning that used a classic single response paradigm have shown that processing conflict problems typically results in lower confidence and longer response latencies (e.g., Botvinick, 2007; De Neys, 2012; Pennycook, Fugelsang, & Koehler, 2015). The question that we want to answer here is whether this is also the case at the initial response stage. Therefore, we contrasted the confidence ratings and response times for the initial response on the conflict problems with those for the initial response on the no-conflict problems. Our central interest here concerns the 11 cases but a full analysis and discussion for each direction of change category is presented in the

[4] We also looked specifically at the "stability" for the 11 responses (i.e., average % of 11 trials among people with at least one 11 trial). This confirmed the overall stability findings with an average across study 1-5 of 66.3%. Stability was slightly lower for the binary 2-response option studies but even here it differed significantly from chance, t(45) = -1427, p < .0001, further arguing against a guessing account. See Table S10 for an overview.

Supplementary Material, section B. In sum, results across our five studies indeed show that 11 responders showed both longer latencies (average increase = 720 ms, SD =113.8 ms, χ^2 (1) = 21.7, $p < 0.0001$, $b = 0.05$) and decreased confidence (average decrease = 12.3 points, SD = 1.9 points, χ^2 (1) = 43.6, $p < 0.0001$, $b = -12.2$) on the conflict vs no-conflict problems. This supports the hypothesis that in addition to the dominant correct intuition the opposing heuristic "10 cents" is also being cued. In other words, System 1 seems to be generating two conflicting intuitions – a logical correct and incorrect heuristic one - in which one is stronger than the other and gets automatically selected as initial response without System 2 deliberation.

Finally, one might also want to contrast the confidence ratings for the different direction of change categories. Previous two-response studies (e.g., Bago & De Neys, 2017; Thompson et al., 2011; Thompson & Johnson, 2014) established that the initial response confidence was lower for responses that got subsequently changed after deliberation (i.e., the "01" and "10" types) than for responses that were not changed (i.e., the "00" and "11" types). It has been suggested that this lower initial confidence (or "Feeling of Rightness" as Thompson et al. refer to it) would be one factor that determines whether reasoners will engage in System 2 deliberation (e.g., Thompson et al., 2011). We therefore looked at the average confidence ratings across Study 1-5. To test the confidence trends, we entered direction of change category and/or response stage (initial or final) as fixed factors in our model (with, as in all our analyses, participants and items as random intercepts). Figure S1 in the supplementary material shows that the pattern reported by Thompson et al. (2011) is replicated in the current study: The initial response confidence for the "01" and "10" categories in which people change their initial response is lower than for responses that are not changed, χ^2 (1) = 193.9, $p < 0.0001$, $b = -27.2$. A similar but less pronounced pattern is observed in the final response stage, χ^2 (1) = 39, $p < 0.0001$, $b = -8.9$. When contrasting the initial and final confidence we also observe that after deliberation there is an overall trend towards increased confidence in the final response stage, χ^2 (1) = 91.6, $p < 0.0001$, $b = 5.3$ (e.g., Shynkaruk & Thompson, 2006; Thompson et al., 2011).

Discussion

Study 2-5 replicated the key finding of our first study. Reasoners who manage to solve the bat-and-ball problem after deliberation often already solved it correctly from the start. This argues against the corrective nature of System 2 deliberation. But this obviously raises a new question. If we don't necessarily need System 2 deliberation to correct our initial intuition, then what do we use or need it for? Why would correct responders ever deliberate, if their intuition already provides them with the correct answer? Here it is important to stress that the fact that people can intuitively generate the correct response, does not imply that the intuitive response will have the exact same characteristics as correct responses that are given after proper deliberation. Even if we accept that System 1 and System 2 might both generate a correct response, the processing characteristics will presumably differ. In other words, there should be some boundary conditions as to what reasoners can do on the basis of mere System 1 processing. Study 6 and 7 focus on this issue.

One of the features that is often associated with System 2 deliberation is that it is *cognitively transparent* (Bonnefon, 2016; Evans & Stanovich, 2013). That is, the output comes "with an awareness of how it was derived" (Bonnefon, 2013). Intuitive processing lacks this explanatory property. Indeed, it is precisely the absence of such processing insight or justification that is often conceived as a defining property of intuitive processing - and one of the reasons to label intuitions as "gut-feelings" (Marewski & Hoffrage, 2015; Mega & Volz, 2014). Bluntly put, this suggest that people might intuitively know and generate a correct answer, but they will not know why it is correct. In Study 6 and 7 we tested this hypothesis by looking at people's response justifications after the initial and final response stage. We hypothesised that although intuitive processes might suffice to estimate the correct answer and produce a correct initial response, people should have little insight into this process and fail to justify why their answer is correct. However, after deliberation in the second

response stage, such proper response justification should become much more likely.

Study 6

Methods

Participants

We recruited 63 Hungarian university students from the Eotvos Lorand University of Budapest (48 female, Mean age = 22.7 years, SD = 1.9 years). These participants received course credit for taking part. In total, 79% of participants reported high school as their highest completed educational level, 21% reported that they already had a post-secondary educational level degree.

Materials and procedure

Since the primary goal of Study 6 (and 7) was to study participant's response justification we made a number of procedural changes to optimize the justification elicitation. Given that explicit justification might be hard (and/or frustrating) we opted to present only half of the Study 1 problems (i.e., two conflict and two no-conflict versions). These items were chosen randomly from the Study 1 problems. Problem content was counterbalanced as in Study 1 and we also used the binary 2-option response format. The procedure followed the same basic two-response paradigm as in Study 1 with the exception that cognitive load was not applied and participants were not requested to enter their response confidence so as to further simplify the task design. The same response deadline as in Study 1 (4 s) was maintained. Note that previous work from our team that contrasted deadline and load treatments indicated that a challenging response deadline may suffice to minimize System 2 engagement in a two-response paradigm (see Bago & De Neys, 2017). After both the initial and final response people were asked the following justification question: "*Could you please try to explain why you selected this answer? Can you briefly justify why you believe it is correct? Please type down your justification below."* There was no time restriction to enter the justification. Whenever participants missed the response deadline for the reasoning problem, they were not presented with the justification question, but rather a message which urged them to make sure to enter their response before the deadline on the next item.

Justification analysis. To analyse participants' justification we defined 8 main justification categories on the basis of an initial screening. Two independent raters categorized the justification responses into these categories. They were in agreement in 86.1% (378 out of 439) of the cases. Cases in which the individual raters did not agree, were afterwards discussed among them with the goal of reaching an agreement (which was reached in all cases). Although our key interest lies in the rate of correct justification, the categorization gives us some insight into the variety of justifications participants spontaneously produce. The eight justification categories along with illustrative examples are presented below. Note that for illustrative purposes we have rephrased the examples into the original bat-and-ball problem units:

Correct math. People referred to the correct mathematical solution (e.g., "because they cost 1.10 together, and the ball costs 1 more than the ball, the ball will be 5 cents and the bat 1.05", "5 cents + 1.05 = 1.10", "110 = x + 100x, 10 = 2x, x = 5", "if the bat is 105 cents and the ball is 5 cents then the bat will be 100 more").

Incorrect math. Participants referred to some sort of mathematical solution but it was not correct (e.g., "1.10 total, so it's got to be 10 cents", "1.10 minus 1 is 10 cents", "because I subtract the given price from the total price, then the rest will be the price of the good in question", "1.10 – 1 = 10") [5].

Unspecified math. Participants referred to mathematical calculation but did not specify the calculation (e.g., "I just did the math", "mental calculation", "this is how the solution comes out with mathematical calculations"; "result of my calculation").

Hunch. People referred to their gut feeling or intuition (e.g., "this is what seemed best, in the

[5] To avoid confusion, note that for the no-conflict control trials, references to the mathematical solution "1.10 - .10 = 1" (i.e., an incorrect math justification on the conflict problems) were obviously scored as correct justifications.

moment of the decision", "this was more sympathetic", "I saw the numbers and based my decision on my intuition", "this automatically came to me as soon as I saw the problem").
Guess. Participants referred to guessing. (e.g., "I guessed", "I could not read it because it was so fast, just clicked on something", "I couldn't really think about the correct solution so I guessed", "was my best guess").
Previous. Participants referred to previous answer without specifying it (e.g., "the justification is the same as before", "my way of thinking is similar to the one I used in the previous task", "I applied same logic as before", "see before").
Other. Any justification that did not fit in other categories (e.g., "I was not sure because I do not have enough time to think it through", " I cannot think and read at the same time", "Hard to tell", " Cannot justify it because I had to answer so quickly that I already forgot").

Exclusion criteria. The same exclusion criteria as in Study 1 were applied: 17.5% of participants were excluded because they had seen the original bat-and-ball problem before and knew the correct response. We further excluded trials where participants failed to provide a response within the deadline (13% of trials). Altogether, 181 trials (out of 208) were further analyzed.

Results and discussion

The accuracy and direction of change pattern in Study 6 was consistent with the pattern we observed in Study 1-5: People typically fail to solve the bat-and-ball problem (37.8% final conflict accuracy), but among correct responders the correct response is frequently generated as initial response (non-correction rate of 59%, see Table 4 and 5 for full details). But clearly, the focus in Study 6 concerns the justifications. We are primarily interested in the proportion of correct justifications for correct conflict responses: In those cases that participants manage to respond correctly, could they also properly justify why their answer was correct? Correct justifications were defined as any minimal reference to the correct mathematical solution (e.g., "because they cost 1.10 together, and the ball costs 1 more, the ball must be 5 cents and the bat 1.05", "5 cents + 1.05 = 1.10", "110 = bat + ball, bat = ball + 100, so ball = 10/2). In these cases we can be sure that participants have some minimal insight into the nature of the correct solution. Table 6 gives an overview of our justification classification for the critical conflict problems (see Supplementary Material Table S5 for the no-conflict problems and Table S9 for the conflict justifications for the individual direction of change categories). The key finding is that correct justifications were indeed much more likely after deliberation than after intuitive responding. Correct justifications on the conflict problems tripled from 20.7% for initial correct responses to 60.6% for final correct responses, χ^2 (1) = 12.4, $p < 0.001$, b = -2.9. This presents some initial support for a boundary condition of correct System 1 responding. Although correct responders might generate the correct solution intuitively, they typically only manage to justify it after deliberation.

However, one might note that our open justification was quite noisy. There were certain types of justifications that were hard to interpret. For example, in the "Unspecified math" category we grouped answers in which participants indicated they "calculated" the response but did not explain how (e.g., "I did the math"). These were most common when participants gave an incorrect response (29% of incorrect cases), but were also observed for correct response (3% of correct cases). Clearly, it is possible that people knew the correct justification, but simply felt there was no need to specify or clarify it. Similarly, participants sometimes also wrote they did "what they did before" (i.e., "Previous" category, 13% of correct responses). Here too it is possible that people could justify the correct solution but did not bother to specify it. To sidestep such interpretational complications, we used a more structured justification elicitation in Study 7. A number of additional methodological improvements also allowed us to further validate the findings.

Study 7

Methods

Participants

A total of 128 Hungarian undergraduates from the Eotvos Lorand University of Budapest were tested (103 female, Mean age = 20.3 years, SD = 1.9 years). Participants received course credit for taking part. 87.5% of participants reported that their highest completed educational level was high school, and 12.5% reported they already had a post-secondary educational level degree.

Materials and procedure

The procedure was based on Study 6 with several methodological changes. The main difference concerned the justification elicitation. To reduce interpretation noise we adopted a semi-structured justification format with four pre-defined answer options that were based on the most frequent justification responses in Study 6. The following illustrates the lay-out:

> *Could you please justify, why do you think that this is the correct response to the question? Please choose from the presented options below:*
>
> ° *"I did the math. Please specify how: ____"*
> ° *"I guessed"*
> ° *"I decided based on intuition/gut feeling"*
> ° *"Other, please specify: ____"*

For the first and fourth answer options participants were also asked to specify their answer. Our rationale was that this format should clarify that we expected them to enter a specification and thereby minimize mere unspecified references to "math/calculations" or "same as before" type responses.

As in Study 6, participants were presented with 2 conflict and 2 no-conflict problems. In Study 7 we adopted further modified content adopted from Trouche (2016; see also Mata et al., 2017). Problems had the same structure as the original bat-and-ball and our Study 6 items but instead of listing the price of two goods, they referred to a different unit (e.g., weight). This should further reduce any possible familiarity effect. Here is an example:

> An apple and an orange weigh 160 grams together. The apple weighs 100 grams more than the orange. How much does the orange weigh?".
>
> ° *60 grams*
> ° *30 grams*

As in our other studies, two problem sets were used in order to counter-balance item content; the conflict items in one set were the control items in the other, and vice-versa. Participants were randomly assigned to one of the sets. A full overview of all problems can be found in the Supplementary Material (section A).

In Study 6 we used the binary, 2-option response format for the reasoning problems. In Study 7 we used both a 2-option and free-response format. Participants were randomly assigned to one of the treatments. The 2-response design was identical to Study 6. The response deadline in the 2-option condition was again 4 s. In the free response condition the deadline was set at 7 s.

In Study 7 we also recorded the time it took for participants to enter their justification. Note that although correct justifications were much more likely after the final response in Study 6, there were still about 20% correct justifications for the initial response on our conflict bat-and-ball problems. One possibility is that some participants used the initial justification stage to start deliberating about their answer. If this is the case, one might expect that the justification response times for correct initial justifications will be affected. The justification latency results in Study 7 lend some support to this hypothesis. Although there were only a handful of correct initial justifications (n = 6), these did tend to take considerably longer (mean = 33.4 s, SD = 2.9 s) than when participants entered an incorrect initial math justification (mean = 17.7, SD = 2.04 s, χ^2 (1) = 3.01, p = 0.08) or a correct initial math justification on no-conflict problems (mean = 17.87 s, SD = 2.9 s, χ^2 (1) = 4.6, p = 0.032, b = 0.25). This suggests that some caution might be needed when interpreting the few correct initial response justifications.

As an additional manipulation check we also included a bogus test question at the end of the survey in Study 7 (Meade & Craig, 2012). Participants might not answer truthfully to our familiarity check ("Have you seen this problem before?") because they feel it is undesirable to answer affirmatively (e.g., because of fears of

being screened out of the study, Wyse, 2013). We included the bogus question *"Have you ever lied in your life*?" to identify such a possible tendency. However, all participants passed this check question and answered affirmative.

Exclusion criteria. The same exclusion criteria as in our other studies were applied. 17.2% of participants were excluded because they had seen the original bat-and-ball problem before and knew the correct response. We further excluded trials where participants failed to provide a response within the deadline (18.5% of trials in the 2-option condition; 12.9% of trials in the free response condition). Altogether, 15.6% of trials were excluded and 358 trials (out of 424) were further analyzed.

Results and discussion

Accuracy and direction of change findings for the 2-option and free response condition in Study 7 can be found in Table 4 and 5. As the tables show, results for the 2-option condition are perfectly in line with the 2-option findings in Study 6 and our other studies: In the vast majority of cases people fail to solve the conflict versions of the bat-and-ball problem, but those who do solve it correctly, often already do so in the initial response phase. The non-correction rate in the 2-option condition reached 62%. However, the pattern in the free response format clearly diverged. Final accuracy reached 48.9% here; the highest rate we observed in any of our studies. The direction of change analysis indicates that this was driven by an extremely high rate of "01" responses (34%) in which the correct response was generated after deliberation. This differed significantly from the 2-option condition rate, χ^2 (1) = 9.1, p = 0.002, b = 1.48. However, the "11" response rate (i.e., correct final responses that are preceded by a correct initial response) did not differ from the 2-option condition, χ^2 (1) = 0.12, p = 0.72, b = -0.48[6], and was in line with what we observed previously. What this suggests is that participants in the free response justification condition showed higher accuracy, not because it was easier to solve the problem intuitively but because it was easier to arrive at the correct response after deliberation. It seems that the combination of being asked to generate your own response and having to justify it boosted deliberation. This boosted deliberation resulted in a non-correction rate of 30%, which is the lowest we observed in any of our studies[7].

Although the positive impact on deliberation is interesting in its own right it does not impact our main justification goal. The key question remains to what extent people can justify their correct initial and final responses: In those cases that participants manage to respond correctly, could they also properly justify why their answer was correct? Table 7 shows the justification results for the conflict problems. We replicate the main finding of Study 6. Both in the 2-option, χ^2 (1) = 54.5, $p < 0.0001$, b = -64.6, and free response format condition, χ^2 (1) = 22.3, $p <$ 0.0001, b = -3.3, correct justifications are much more likely for correct final than for correct initial responses. With the structured justification elicitation in Study 7 we obtained correct justification in over 90% of final correct responses (2-option: 96.2%; free response: 90.7%). This directly establishes that whenever people give a correct response after deliberation, they have little trouble to justify it. However, such justification is much rarer for correct initial responses (2-option: 9.1%; free response: 26.7%). A closer look at Table 7 shows that the dominant justifications for correct initial responses were references to intuition and guessing. These types of justifications were completely absent for correct final responses. Taken together, the results of our justification studies provide clear evidence for a boundary condition of correct System 1 intuitions. We can estimate the correct answer intuitively, but we don't know how we do it. Our System 1

[6] The random effect of items was left out of the model because of convergence problems.

[7] For completeness, we also looked at response accuracy on the first conflict problem in our justification studies. Across studies 6 and 7 the non-correction rate was 33.3% (individual studies range from 23.3% to 60% non-correction; average 2-response format = 43.3%, free response = 23.3%, see supplementary Table S8). Although these data concern a limited number of observations they further indicate that correct intuitive responding is observed from the start of the experiment, as we observed in Study 1-5.

knowledge is not cognitively transparent (Bonnefon, 2016).

General Discussion

Influential work in the reasoning and decision making field since the 1960s has popularized a corrective view of human reasoning (Evans & Stanovich, 2013; Kahneman, 2011). This view entails that sound thinking in reasoning tasks such as the bat-and-ball problem often requires correction of fast, intuitive thought processes by slower and more demanding deliberation. The present study questions this idea. We focused on the very problem that has been widely featured as the paradigmatic illustration of the corrective view, the bat-and-ball problem (e.g., Frederick, 2005; Kahneman, 2011). By adopting a two response paradigm in which people were required to give an initial response under time-pressure and cognitive load we aimed to identify the intuitively generated response that preceded the final response given after deliberation. Across our studies we consistently observed that correct final responses are often non-corrective in nature. In a substantial number of cases, reasoners who manage to answer the bat-and-ball problem correctly after deliberation already solved it correctly when they reasoned under conditions that force reliance on intuition in the initial response phase. In other words, people who solve the bat-and-ball problem do not necessarily need to deliberate to correct their intuitions, their intuitions are often already correct.

These findings point to at least two fundamental implications about the way we conceive intuitive and deliberate thinking or System 1 and 2. On one hand, it suggests that we might need to upgrade our view of the intuitive System 1. Although System 1 can frequently cue incorrect intuitions, it also generates correct intuitions. Among correct responders it are these correct intuitions that will often dominate. Consequently, even when we're faced with the notorious bat-and-ball problem, intuitive thinking is less ignorant or "smarter" than traditionally assumed. On the other hand, the upgrading of System 1 also suggests we need to revise the role of System 2. When the correct response can be generated intuitively, the central role of System 2 deliberation cannot exclusively lie in a correction process. The results of our justification studies suggest that instead of in correction, the contribution of deliberate processing in these cases might rather lie in its cognitive transparency (Bonnefon, 2016). We observed that whereas people don't manage to explain why their initial response is correct, they seem to have little difficulties in providing such correct justifications after deliberation. Clearly, being able to produce a proper justification for one's insights is quite crucial. This was well-understood by the leading scientists we cited to illustrate the "intuition-as-a-guide" view: Although Einstein and Poincaré wanted to highlight the key role of intuitive processes, they also stressed the importance of subsequent deliberation. Bluntly put, Kukulé and Newton would not have managed to convince their peers, if they had simply claimed their ideas were correct because they "felt it". Hence, even among the historical proponents of the key role of intuitive thinking for sound reasoning there was never any question that intuitive insight will need further reflection and validation to be fully developed. In other words, the initial intuitive insight is important but does not suffice. What the present study suggests is that this view on intuitive reasoning in which deliberation is helping to validate an initial intuitive insight might be a more appropriate model to conceive of human reasoning than a view in which the key function of deliberation merely lies in correction of erroneous intuitions.

In recent years there have been a number of popular accounts that have celebrated the advantages of intuitive thinking over deliberate thinking (Dijksterhuis, 2011; Gigerenzer, 2007; Gladwell, 2005). Against this backdrop it should be stressed that our call to upgrade the role of System 1 and our arguments against the mere corrective view of System 2 should not be conceived as a claim to downgrade the importance of System 2 deliberation. First, across all our studies there were always instances in which System 2 correction did occur (i.e. "01" cases). Hence, the prototypical corrective pattern in which an initially faulty intuition is corrected after deliberation is also observed. Second, as we alluded to above, the fact that deliberation does not necessarily play a role in correction does not

imply it is not important for other reasons. Our findings suggest that one such reason might be its cognitive transparency and the fact that after deliberation people are able to come up with a proper justification. Hence, deliberation can help to produce a good explanation or argument for why a response is correct. Such arguments are critical for communicative purposes (e.g., Mercier & Sperber, 2011). What is true for scientific discussions is also true for daily life: we will not be very successful in convincing others that our answer to a problem is correct, if we can only tell them that we feel it is right. If we come up with a good explanation, however, people will be much more likely to change their mind (Trouche, Sander, & Mercier, 2014). Such argumentative persuasion has been argued to be the evolutionary driving force behind the development of the human capacity to reason (Mercier & Sperber, 2011). Indeed, the human success as a social and cultural species is hard to imagine without an ability to communicate and transmit good problem solutions[8]. Hence, it would be foolish to interpret our findings and arguments against the corrective view of deliberation or System 2 as evidence against the role of deliberation in thinking per se.

In addition, one needs to bear in mind that although our findings present evidence for the possible non-corrective nature of correct responding, most people are still biased and fail to give the correct answer when solving the bat-and-ball problem. In absolute numbers, incorrect "10 cents" responses are still much more common than correct "5 cents" responses. Solving the bat-and-ball problem correctly is still exceptional. The key issue is that in those cases it does occur, the correct response is often already generated intuitively. But in absolute terms such correct intuitive response generation remains rare. Obviously, the point is not that System 1 is always correct, the point is that it can be correct and is often already so for reasoners who respond correctly after having deliberated.

Likewise, our findings should not be taken to imply that System 2 deliberation *cannot* be used to correct faulty intuitions or that such correction is never required. The key assumption we test in the present study is whether correct responding results from deliberate correction of a faulty intuition. We examined whether sound reasoners respond correctly precisely because they manage to complete this correction process. Our results show that this is not necessarily the case. Often, correct responders have nothing to correct. However, this does not imply that correction is redundant for everyone. Our results do not imply that everyone manages to generate the correct answer intuitively. As we noted, the vast majority of reasoners gives the faulty intuitive "10 cents" response both at the initial response stage and after deliberation. Hence, not everyone will generate a correct (intuitive) response. For most reasoners, the incorrect intuition will dominate. Consequently, our empirical results directly argue against the idea that correction of System 1 is never needed. And it might very well be the case that additional deliberation could be helpful in these cases. Imagine that we devise an intervention procedure that allows us to train biased reasoners to deliberately correct. This might very well reduce bias and boost correct responses. Our results do not speak to this issue. Hence, the present findings do not imply that interventions are pointless or that deliberate correction is impossible or redundant. Our point here is that spontaneous sound reasoning does not necessarily require such correction. It is this central assumption of the corrective view that our results question.

As we clarified in the introduction, literally hundreds of studies have focused on the bat-and-ball problem in the last ten years. One common objection to our study might be that if the non-correction phenomenon and correct "5 cents" intuitions are really so ubiquitous, then why has this phenomenon not been documented previously? We believe that the simple answer is that scholars haven't really looked for it. Note

[8] Clearly, deliberation might have further additional benefits beyond communication per se. For example, another value of deliberate explanation might lie in the improvement of one's own understanding which can facilitate knowledge transfer to other relevant problems (e.g., Wertheimer, 1945). Obviously, this does not preclude – as our justification data shows - that deliberation can also be used to generate incorrect justifications for incorrect responses, a process often referred to as "rationalization" (e.g., Wason & Evans, 1975; Pennycook et al., 2015). Our claim here concerns the reasoning of correct responders.

that we were only able to identify the correct intuitions by carving up the reasoning process with the two-response paradigm (Thompson et al., 2011). In a traditional "one-response" experiment System 1 and 2 processing will go hand in hand. That is, the correct intuitive responders will typically back up their System 1 processing with System 2 deliberation to validate their answer. Reasoners will not end their reasoning process after they have come up with a correct intuitive response. This implies, for example, that the final response generation for those who give the correct response will still take longer than for those who give an incorrect response and do not engage (or engage less profoundly) in System 2 deliberation[9]. It is only by experimentally isolating the initial reasoning stage that we were able to demonstrate the correct nature of the initially generated response in these cases. Bluntly put, it is unlikely that a pure correct intuitive response will be observed "in the wild". Just as with other non-naturalistically perceivable scientific phenomena (Niaz, 2009) we suspect that this helps to explain why the non-corrective nature of System 2 deliberation has gone largely unnoticed in empirical studies.

To be very clear, we are not the first to point towards the potential of intuitive processing. We referred to the intuition-as-guide view to illustrate how more than a century ago leading scientists already argued that the origin of their key insights relied on intuitive processing. Furthermore, within the cognitive sciences various scholars have developed related ideas in a range of frameworks (e.g., Dijksterhuis' Unconscious Thinking Theory, 2011; Gigerenzer's Fast and Frugal Heuristics, 2007; Klein's Naturalistic Decision Making, 2004; Reyna's Fuzzy-Trace Theory, 2012). More specifically, Peters (2012) has explicitly raised the possibility that good reasoners might manage to arrive at the correct response in the bat-and-ball problem precisely because they have correct intuitions. The critical contribution of our study lies in the empirical demonstration of this phenomenon. More generally, one might argue that even the traditional dual process framework can accommodate the present findings with some additional qualification. One general feature of dual process models is that with practice and experience processes that initially need System 2 deliberation can be automatized and handled by System 1 (Epstein, 1994; Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996). In a way, such automatization is precisely what we hope to achieve in many teaching or learning contexts (e.g., Schneider & Shiffrin, 1977). Solving the bat-and-ball problem boils down to solving the algebraic equation "X + Y = 1.10, Y = 1 + X, Solve for X". This is something that all educated adults have done at length in their high school math classes. One can account for the present findings by assuming that years of exposure to this type of problem solving helped sound reasoners to automatize the solution process. Consequently, there would no longer be a need for deliberate correction. We would not object to the idea that such automatization may lie at the heart of the currently observed correct intuitive responding. Hence, although the current findings argue against the traditional corrective dual process view they do not necessarily invalidate the wider framework itself. But it underscores the need for any viable dual process model to fully recognize and embrace the potential of System 1 (e.g., Stanovich, 2018; Thompson, Pennycook, Trippas, & Evans, 2018, for related suggestions).

Proponents of the traditional corrective view can try to point to a fundamental methodological limitation of our study. We used a two-response paradigm in which we tried to make sure that the initial responses were intuitive in nature by combing an instruction, time-pressure, and load manipulation. All these manipulations have been previously shown to limit System 2 deliberation. By combining them we believe we created one the most stringent and purest operationalisations of System 1 processing that have been adopted in the dual process literature to date. However, critics might argue

[9] This also implies that one's performance on the bat-and-ball or related problems (e.g., items from the Cognitive Reflection test, Frederick, 2005) is still a valid measure of one's tendency to reflect or deliberate. The present data indicate that correct responders are still more likely to deliberate (or are better at deliberation) than incorrect responders (i.e., after deliberation they manage to justify their response). The point is simply that the nature of this deliberation process does not necessarily lie in a correction process but rather in a justification process.

that we can never be completely sure that we eliminated all System 2 processing. In theory, this is correct. The general problem is that dual process theories are underspecified (Kruglanski, 2013). The framework often entails that System 2 is slower and more demanding than System 1 but gives us no unequivocal a priori criterion that allows us to classify a process as System 1 or 2 (e.g., takes at least x time, or x amount of load). Consequently, as long as we keep on observing correct initial responses, one can always argue that these will disappear "with just a little bit more load/time pressure". Note, however, that the corrective assumption becomes unfalsifiable at this point. Any negative evidence can always be explained away by arguing that the procedure did not fully rule out deliberation.

On the other hand, we readily agree that further testing and replication is always welcome. The present study is but the first to document the non-corrective nature of the "5 cents" response. This finding questions the common wisdom in the field. As one of our reviewers put it, "extraordinary claims need extraordinary evidence". Although we ran seven studies to test the robustness of our findings, further validation remains important. For example, future studies might try to identify the precise boundary conditions under which correct initial responses are observed by systematically setting stricter deadlines and/or adopting alternative, more challenging load tasks, testing for individual differences, etc. In this light, one might also note that we used an operational definition of intuitive, System 1 processing that focused on speed and effort. Although these are typical features that are often used to differentiate System 1 and 2 processing, other characteristics can be put forward. For example, one such alternative defining characteristic of System 1 processing might be its "autonomy" (i.e., processing is either mandatory given the presence of triggering conditions – System 1 – or not mandatory – System 2, e.g., Pennycook, 2017). In as far as autonomy is independent from speed and effort, one could still argue that the initial response in our paradigm results from deliberate, System 2 processing (i.e., the fast and undemanding initial response would not be mandatory and hence, be deliberate and not intuitive in nature). We personally have some difficulties envisaging how one would operationalize and test such an account but we acknowledge that it is a theoretical possibility. Another limitation is that although we identified correct initial responses we haven't specified the nature of these responses. The precise mechanism behind the generation of correct "5 cents" intuitions is not clear. Above we alluded to the possibility that these might result from an automatization process through repeated exposure in a formal educational setting. Obviously, this is a speculative claim that we haven't tested directly. We fully agree that pinpointing the precise nature of the postulated intuitions remains an important challenge (De Neys, 2017).

A final objection against the current work might be that it focused exclusively on the bat-and-ball problem. One easy way out for proponents of the corrective view would be to argue that our findings simply imply that the field has mistakenly characterized the bat-and-ball problem as a prototypical example of the correction process. Hence, the corrective view could be maintained but it would simply need to change its poster boy. This is problematic for several reasons. First, if we post hoc classify a particular task that does not fit the predictions as an exceptional case, we end up with a framework that has hardly any explanatory power. But more critically, the current findings have also been observed with other classic reasoning tasks such as belief-bias syllogisms and base-rate neglect tasks (Bago & De Neys, 2017; Newman et al., 2017). Hence, it is not the case that the observed non-correction is some idiosyncratic peculiarity of the bat-and-ball problem that would fail to generalize to other tasks. Nevertheless, belief-bias and base-rate neglect problems are easier (i.e., show lower bias rates) than the bat-and-ball problem and there is less a priori agreement on how representative they are to test the corrective view (Aczel et al., 2016; Evans, 2017; Mata et al., 2017; Singmann et al., 2014; Travers et al., 2016; Pennycook et al., 2012). By showing that the corrective prediction does not hold up in the specific case that is considered to be one of its strongholds, we believe we provide a critical test that should at least force us to question a strict corrective dual process view of deliberation. Deliberation is undoubtedly critical for human

thinking, but sound reasoners do not necessarily need it to correct faulty intuitions.

Acknowledgements

Bence Bago is supported by a fellowship from the Ecole des Neurosciences de Paris Ile-de-France. This research was also supported by a research grant (DIAGNOR, ANR-16-CE28-0010-01) from the Agence National de la Recherche. We like to thank Jean-François Bonnefon for valuable feedback on an earlier version of this manuscript and Balazs Aczel for helping us recruiting the Hungarian participants. All raw data can be retrieved from osf.io/ycjd9.

References

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

Alós-Ferrer, C., Garagnani, M., & Hügelschäfer, S. (2016). Cognitive reflection, decision biases, and response times. *Frontiers in Psychology*, *7*.

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, *158*, 90–109.

Bonnefon, J.-F. (2013). New ambitions for a new paradigm: Putting the psychology of reasoning at the service of humanity. *Thinking & Reasoning*, *19*(3–4), 381–398.

Bonnefon, J.-F. (2016). The Pros and Cons of Identifying Critical Thinking with System 2 Processing. *Topoi*, 1–7.

Botvinick, M. M. (2007). Conflict monitoring and decision making: reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(4), 356–366.

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. Blaisdell, L. Huber, & A. Young (Eds.), *Rational animals, irrational humans* (pp. 235–252).

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130.

De Neys, W. (2012). Bias and conflict a case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38.

De Neys, W. (2017). Bias, conflict, and fast logic : Towards a hybrid dual process future ?. In W. De Neys (Ed.), Dual Process Theory 2.0 (pp. 47-65). Oxon, UK: Routledge.

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.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load. *Experimental Psychology (Formerly Zeitschrift Für Experimentelle Psychologie)*, *54*(2), 128–133.

De Neys, W., & Verschueren, N. (2006). Working memory capacity and a notorious brain teaser: The case of the Monty Hall Dilemma. *Experimental Psychology*, *53*(2), 123–131.

Dijksterhuis, A. P. (2011). *Het slimme onbewuste*. Prometheus.

Ellenberg, J. (2015). *How not to be wrong: The power of mathematical thinking*. Penguin.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, *49*(8), 709–724.

Evans, J. St. B. T. (2010). *Thinking Twice: Two Minds in One Brain*. Oxford: Oxford University Press.

Evans, J. St. B.T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241.

Evans, J. St. B. T. (2017). Dual Process Theories: Perspectives and problems. In W. De Neys (Ed.), *Dual Process Theory 2.0* (pp. 137–155). Oxon, UK: Routledge.

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning*, *15*(2), 105–128.

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*(4), 25–42.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. London: Penguin Books.

Gilhooly, K. J. (2016). Incubation and intuition in creative problem solving. *Frontiers in psychology*, *7*, 1076.

Gladwell, M. (2005). Blink: The guide to thinking without thinking. *Little, Brownand Company, New York*.

Haigh, M. (2016). Has the Standard Cognitive Reflection Test Become a Victim of Its Own Success? *Advances in Cognitive Psychology*, *12*(3), 145–149.

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.

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56–64.

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

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, *49*, 81.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. *The Cambridge handbook of thinking and reasoning*, 267-293.

Klein, G. A. (2004). *The power of intuition: How to use your gut feelings to make better decisions at work*. Crown Business.

Kruglanski, A. W. (2013). Only one? The default interventionist perspective as a unimodel—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*(3), 242–247.

Lehrer, J. (2011). The Science of Irrationality. Retrieved from https://www.wsj.com/articles/SB10001424052970203633104576625071820638808

Levitt, S. D., & Dubner, S. J. (2010). *Freakonomics* (Vol. 61). Sperling & Kupfer editori.

Marewski, J. N., & Hoffrage, U. (2015). Modeling and aiding intuition in organizational decision making. *Journal of Applied Research in Memory and Cognition*, *4*, 145–311.

Mastrogiorgio, A., & Petracca, E. (2014). Numerals as triggers of System 1 and System 2 in the 'bat and ball'problem. *Mind & Society*, *13*(1), 135–148.

Mata, A., & Almeida, T. (2014). Using metacognitive cues to infer others' thinking. *Judgment and Decision Making*, *9*(4), 349–359.

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: an attentional account of reasoning errors. *Psychonomic Bulletin & Review*, 1–7.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455.

Mega, L. F., & Volz, K. G. (2014). Thinking about thinking: implications of the introspective error for default-interventionist type models of dual processes. *Frontiers in Psychology*, *5*.

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in cognitive sciences*, *22*, 280-293.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(02), 57–74.

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.

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in cognitive sciences, 14*, 435-440.

Newman, I., 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.

Niaz, M. (2009). *Critical appraisal of physical science as a human enterprise: Dynamics of scientific progress* (Vol. 36). Springer Science & Business Media.

Oesper, R. E. (1975). *The human side of scientists*. University Publications, University of Cincinnati.

Pennycook, G. (2017). A perspective on the theoretical foundation of dual process models. In W. De Neys (Ed.), Dual Process Theory 2.0. Oxon, UK: Routledge.

Pennycook, G., De Neys, W., Evans, J. St. B.T., Stanovich, K. E., Thompson, V. (2018). The mythical dual-process typology. Manuscript submitted for publication.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, *124*(1), 101–106.

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.

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review*, *19*(3), 528–534.

Peters, E. (2012). Beyond comprehension the role of numeracy in judgments and decisions. *Current Directions in Psychological Science*, *21*(1), 31–35.

Poincaré, H. (1914). Science and Method, translated by F. *Maitland, Preface by B. Russell (Thomas Nelson and Sons, London, 1914)*.

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision Making*, *7*(3), 332–359.

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122*(2), 166–183.

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition, 34*(3), 619-632.

Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, *6*, 532.

Singmann, H., Klauer, K. C., & Kellen, D. (2014). Intuitive logic revisited: new data and a Bayesian mixed model meta-analysis. *PloS One*, *9*(4), e94223.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10393–10398.

Stanovich, K. E. (2018). Miserliness in human cognition: the interaction of detection, override and mindware. *Thinking & Reasoning*.

Stieger, S., & Reips, U.-D. (2016). A limitation of the Cognitive Reflection Test: familiarity. *PeerJ*, *4*, e2395.

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

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: exploring the ways individuals solve the test. *Thinking & Reasoning*, 1–28.

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244.

Thompson, V. A., Pennycook, G., Trippas, D., & Evans, J. St. B. T. (2018). Do smart people have better intuitions. *Journal of Experimental Psychology: General*.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140.

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. *Cognition*, *150*, 109–118.

Trouche, E. (2016). *Le raisonnement comme competence sociale: Une comparaison expérimentale avec les théories intellectualistes*. Unpublished PhD thesis.

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.

Wyse, S. E. (2013). Don’t disappoint your online survey respondents. Retrieved from https://www.snapsurveys.com/blog/disappoint-online-survey-respondents/

Table 1. Initial and final average (SD) response accuracy in study 1-5.

Study	Response format	Conflict items		No-conflict control items	
		Initial response	Final response	Initial response	Final response
Study1	2 response	20.1% (40.2)	24.5% (43.1)	93.4% (24.9)	96.0% (19.6)
Study2a	2 response	7.7% (26.7)	7.3% (26.1)	94.6% (22.6)	96.5% (18.5)
Study2b	4 response	9.0% (28.6)	10.8% (31.1)	94.9% (22)	95.1% (21.6)
Study3a	2 response	23.9% (42.8)	28.3% (45.2)	96.3% (18.9)	97.1% (17)
Study3b	4 response	19.0% (39.4)	29.5% (45.3)	95.6% (20.6)	97.1% (17)
Study4	Free response	10.9% (31.3)	13.1% (33.9)	96.4% (18.8)	99.3% (8.5)
Study5	Free response	30.5% (46.2)	41.4% (49.4)	93.8% (24.2)	98.1% (13.6)
Average	2 response	14.7% (34.4)	15.4% (36.1)	94.5% (22.9)	96.4% (18.6)
	4 response	11.1% (31.5)	14.6% (35.4)	95.0% (21.7)	95.5% (20.7)
	Free response	20.4% (40.4)	26.8% (44.4)	95.0% (21.9)	98.7% (11.5)
Overall average		13.8% (34.5)	16.8% (37.4)	94.7% (22.3)	96.5% (18.5)

Table 2. Frequency of direction of change categories in study 1-5 for conflict items. Raw number of trials are in brackets.

Study	Response format	Direction of change category				Non-correction (11/11+01)
		11	00	10	01	
Study1	2 response	16.5% (45)	71.8% (196)	3.7% (10)	8.1% (22)	67.2%
Study2a	2 response	4.5% (24)	89.5% (475)	3.2% (17)	2.8% (15)	61.5%
Study2b	4 response	8.6% (42)	88.8% (436)	0.4% (2)	2.2% (11)	79.2%
Study3a	2 response	16.7% (23)	64.5% (89)	7.2% (10)	11.6% (16)	59%
Study3b	4 response	19.0% (26)	71.5% (98)	-	9.5% (13)	66.7%
Study4	Free response	10.9% (15)	86.7% (119)	-	2.2% (3)	83.3%
Study5	Free response	29.7% (38)	57.8% (74)	0.8% (1)	11.7% (15)	71.7%
Average	2 response	9.8% (92)	80.7% (760)	3.9% (37)	5.6% (53)	63.4%
	4 response	10.8% (68)	85.0% (534)	0.3% (2)	3.8% (24)	73.9%
	Free response	20.0% (53)	72.8% (193)	0.4% (1)	7.0% (18)	74.6%
Overall average		10.9% (198)	81.7% (1487)	2.2% (40)	5.2% (95)	67.6%

Table 3. Frequency of stability index values on conflict items in Study 1-5. The raw number of participants for each value is presented between brackets.

Study	Response format	Stability index value					Average stability
		<33%	50%	66%	75%	100%	
Study1	2 response	3.6% (3)	10.7% (9)	8.3% (7)	9.5% (8)	67.9% (57)	87.1%
Study2a	2 response	2.4% (4)	3.0% (5)	4.8% (8)	6.6% (11)	83.2% (139)	93.7%
Study2b	4 response	3.5% (5)	1.4% (2)	2.8% (4)	4.2% (6)	88.1% (126)	95.0%
Study3a	2 response	4.9% (2)	19.5% (8)	7.3% (3)	12.2% (5)	56.1% (23)	81.5%
Study3b	4 response	2.3% (1)	7.0% (3)	14.0% (6)	2.3% (1)	74.4% (32)	89.7%
Study4	Free response	-	2.6% (1)	5.1% (2)	-	92.3% (36)	97.0%
Study5	Free response	-	19.5% (8)	9.8% (4)	9.8% (4)	61.0% (25)	84.6%
Average	2 response	3.1% (9)	7.5% (22)	6.2% (18)	8.2% (24)	75.0% (219)	90.0%
	4 response	3.2% (6)	2.7% (5)	5.4% (10)	3.8% (7)	84.9% (158)	93.7%
	Free response	-	11.3% (9)	7.5% (6)	5.0% (4)	76.5% (61)	90.6%
Overall average		2.7% (15)	6.5% (36)	6.1% (34)	6.3% (35)	78.5% (438)	91.3%

Table 4. Initial and final accuracies (SD) in justification studies (Study 6-7).

Study	Response format	Conflict items		No-conflict control items	
		Initial response	Final response	Initial response	Final response
Study 6	2 response	32.2% (32.2)	37.8% (37.8)	86.8% (86.8)	94.5% (94.5)
Study 7a	2 response	28.9% (45.7)	34.2% (47.8)	90.8% (29.1)	100% (0)
Study 7b	Free response	17% (37.8)	48.9% (50.3)	92.5% (26.4)	100% (0)

Table 5. Frequency of each direction of change category for the conflict items in each justification study (Study 6-7). Raw number of trials are in brackets.

Study	Response format	Direction of change category				Non-correction (11/11+01)
		11	00	10	01	
Study 6	2 response	22.2% (20)	52.2% (47)	10% (9)	15.5% (14)	58.8%
Study 7a	2 response	21.1% (16)	57.9% (44)	7.9% (6)	13.2% (10)	61.5%
Study 7b	Free response	14.77% (13)	48.86% (43)	2.27% (2)	34.1% (30)	30.2%

Table 6. Frequency of different types of justifications for conflict items in Study 6 and 7 (raw number of justifications in brackets).

Study	Justification	Initial response		Final response	
		Correct	Incorrect	Correct	Incorrect
Study 6	Correct math	20.7% (6)	1.6% (1)	60.6% (20)	5.6% (3)
2 response	Incorrect math	-	24.6% (15)	-	33.3% (18)
	Unspecified math	-	14.8% (9)	6.1% (2)	44.4% (24)
	Hunch	3.4% (1)	6.6% (4)	6.1% (2)	-
	Guess	34.5% (10)	9.8% (6)	3% (1)	3.7% (2)
	Previous	20.7% (6)	6.6% (4)	6.1% (2)	1.9% (1)
	Other	20.7% (6)	36.1% (22)	18.2% (6)	11.1% (6)
Study 7	Correct math	9.1% (2)	1.9% (1)	96.2% (25)	2% (1)
2 response	Incorrect math	-	37% (20)	-	69.4% (34)
	Unspecified math	9.1% (2)	9.3% (5)	3.8% (1)	16.3% (8)
	Hunch	45.5% (10)	27.8% (15)	-	2% (1)
	Guess	36.4% (8)	24.1% (13)	-	10.2% (5)
	Other	-	-	-	-
Study 7	Correct math	26.7%(4)	-	90.7% (39)	2.2% (1)
Free response	Incorrect math	-	38.6% (28)	-	68.9% (31)
	Unspecified math	6.7% (1)	4.1% (3)	4.7% (2)	17.8% (8)
	Hunch	46.7% (7)	30.1% (22)	-	4.4% (2)
	Guess	13.3% (2)	23.3% (17)	-	6.7% (3)
	Other	6.7% (1)	4.1% (3)	4.7% (2)	-

Supplementary Material

A. Problems used in study 1-7

Items used in Study 1-6:

	Conflict version	Control version
1	A pencil and an eraser cost $1.10 in total. The pencil costs $1 more than the eraser. How much does the eraser cost?	A pencil and an eraser cost $1.10 in total. The pencil costs $1. How much does the eraser cost?
2	A magazine and a banana cost $2.60 in total. The magazine costs $2 more than the banana. How much does the banana cost?	A magazine and a banana cost $2.60 in total. The magazine costs $2. How much does the banana cost?
3	A cheese and a bread cost $2.90 in total. The cheese costs $2 more than the bread. How much does the bread cost?	A cheese and a bread cost $2.90 in total. The cheese costs $2. How much does the bread cost?
4	An apple and an orange cost $1.80 in total. The apple costs $1 more than the orange. How much does the orange cost?	An apple and an orange cost $1.80 in total. The apple costs $1. How much does the orange cost?
5	A sandwich and a soda cost $2.50 in total. The sandwich costs $2 more than the soda. How much does the soda cost?	A sandwich and a soda cost $2.50 in total. The sandwich costs $2. How much does the soda cost?
6	A hat and a ribbon cost $4.20 in total. The hat costs $4 more than the ribbon. How much does the ribbon cost?	A hat and a ribbon cost $4.20 in total. The hat costs $4. How much does the ribbon cost?
7	A coffee and a cookie cost $2.40 in total. The coffee costs $2 more than the cookie. How much does the cookie cost?	A coffee and a cookie cost $2.40 in total. The coffee costs $2. How much does the cookie cost?
8	A book and a bookmark cost $3.30 in total. The book costs $3 more than the bookmark. How much does the bookmark cost?	A book and a bookmark cost $3.30 in total. The book costs $3. How much does the bookmark cost?

Response options for each of the problems in Study 1-6:

	2 response options	4 response options
1	5, 10	1, 5, 10, 15
2	30, 60	15, 30, 60, 90
3	45, 90	15, 45, 90, 135
4	40, 80	20, 40, 80, 120
5	25, 50	5, 25, 50, 75
6	10, 20	5, 10, 20, 30
7	20, 40	10, 20, 40, 60
8	15, 30	5, 15, 30, 45

Items used in Study 7:

	Conflict version	Control version
1	An apple and an orange weigh 160 grams altogether. The apple weighs 100 grams more than the orange. How much does the orange weigh?	An apple and an orange weigh 160 grams altogether. The apple weighs 100 grams. How much does the orange weigh?
2	In a shop there are 250 PCs and MACs altogether. There are 200 more PCs than MACs. How many MACs are there in the shop?	In a shop there are 250 PCs and MACs altogether. There are 200 PCs. How many MACs are there in the shop?
3	Altogether, a book and a magazine have 330 pages. The book has 300 pages more than the magazine. How many pages does the magazine have?	Altogether, a book and a magazine have 330 pages. The book has 300 pages. How many pages does the magazine have?
4	In total, a plumber and an electrician work 240 days. The electrician works 200 days more than the plumber. How many days does the plumber work?	In total, a plumber and an electrician work 240 days. The electrician works 200 days. How many days does the plumber work?

B. Conflict detection analysis Study 1-5

For each direction of change category one may ask whether reasoners are faced with two competing intuitions at the first response stage. We can address this question by looking at the contrast between conflict and control problems. If conflict problems cue two conflicting initial intuitive responses, people should process the problems differently than the no-conflict problems (in which such conflict is absent) in the initial response stage and show lower confidence and longer response latencies (e.g., Botvinick, 2007; De Neys, 2012; Johnson et al., 2016; Pennycook et al., 2015) when solving the conflict problems. Therefore, we contrasted the confidence ratings and response times[10] for the initial response on the conflict problems with those for the initial response on the no-conflict problems for each of the four direction of change categories. Note that we used only the dominant control 11 category for this contrast (which we will refer to as "baseline"), as responses in the other control direction of change categories cannot be interpreted unequivocally.

Table S1 (confidence) and S2 (latencies) show the results. Visual inspection of Table S1 indicates that there is a general trend towards a decreased initial confidence when solving conflict problems for all direction of change categories. However, this effect is much larger for the 01 and 10 cases in which reasoners subsequently changed their initial response. This suggests that although reasoners might be experiencing some conflict between competing intuitions in all cases, this conflict is much more pronounced in the 10 and 01 case. Latency data in Table S2 mirrors this pattern. In all change categories, it took more time to give a response on conflict items but this latency increase is most pronounced in case people ended up changing their initial response.

To analyse the data statistically we again created mixed effect multi-level models (Baayen, Davidson, & Bates, 2008; Kuznetsova, Brockhoff, & Christensen, 2017). We ran a separate analysis for each of the four direction of change conflict problem categories and we analysed both confidence and reaction times. In the analysis, the confidence or reaction time for the initial response in a given direction of change category in question was contrasted with the initial response confidence or reaction time for 11 control problems which served as our baseline. We will refer to this contrast as the conflict factor. The conflict factor was entered as fixed factor, and participants and items were entered as random factor. We also entered the response format (2 response vs 4 response vs free response) as fixed factor mainly to test for an interaction with the conflict factor. In the cases in which we found a significant interaction we also analysed each response format condition separately. We were not interested in main effects of the response format (e.g., simply because of the different deadlines, responses will be faster for some studies than for others) and did not analyse it further to avoid spurious findings. Note that prior to analysis reaction times were log-transformed to normalize the distribution, and analysis were performed on the log-transformed data.

11 Category. In terms of confidence, we found that conflict improved model fit significantly, χ^2 (1) = 43.6, $p < 0.0001$, as well as the main effect of response format, χ^2 (3) = 7.63, $p = 0.022$, but not their interaction, χ^2 (5) = 1.3, $p = 0.52$. Hence, people were less confident in the 11 conflict category than in the baseline, b = -12.2, t (193.2) = -10.83, $p <$

[10] Note that the initial response time analysis should be interpreted with some caution. Previous two-response studies established that response times do not reliably track conflict detection effects reflected in confidence ratings at the initial response stage (Bago & De Neys, 2017; Thompson & Johnson, 2014).

0.0001. Similar results were found with regard to reaction times as well; conflict improved model fit significantly, χ^2 (1) = 21.73, $p < 0.0001$, as well as the main effect of response format, χ^2 (3) = 554.9, $p < 0.0001$, but not their interaction, χ^2 (5) = 2.23, $p = 0.33$. Thus, it took more time to give a response in the 11 conflict category, than in the baseline, $b = 0.05$, t (127.8) = 5.4, $p < 0.0001$.

00 category. With regard to confidence, the main effect of conflict improved model fit significantly, χ^2 (1) = 12.6, $p = 0.0004$, but neither response format, χ^2 (3) = 5.5, $p = 0.06$, nor their interaction did, χ^2 (5) = 4.6, $p = 0.1$. Hence, people were less confident in their response in the 00 conflict category than in baseline, $b = -3.1$, t (12.93) = -4.3, $p = 0.0008$. For reaction times, we found that conflict improved model fit, χ^2 (1) = 7.04, $p = 0.008$, along with the main effect of response format, χ^2 (3) = 559.3, $p < 0.0001$, but not their interaction, χ^2 (5) = 1.84, $p = 0.4$. Hence, it took people more time to give a response in the 00 conflict category than in the baseline, $b = 0.02$, t (13.7) = 2.9, $p = 0.01$.

10 category. For confidence, we found that model fit was improved by conflict, χ^2 (1) = 50.1, $p < 0.0001$, and the main effect of response format, χ^2 (3) = 6.6, $p = 0.036$. Their interaction did not improve model fit significantly, χ^2 (5) = 3.3, $p = 0.19$. Therefore, people were less confident in the 10 conflict category than in the baseline, $b = -49.5$, t (834.9) = -21.9, $p < 0.0001$. For reaction times, we found that only response format improved the model fit significantly, χ^2 (3) = 557.3, $p < 0.0001$, but not conflict, χ^2 (1) = 1.25, $p = 0.26$, and not their interaction, χ^2 (5) = 1.9, $p = 0.39$.

01 category. Regarding confidence, we found that conflict improved model fit significantly, χ^2 (1) = 50.8, $p < 0.0001$. There was no main effect of response format, χ^2 (3) = 3.4, $p = 0.18$, but format and conflict did interact, χ^2 (5) = 60.25, $p < 0.0001$. We analysed each of the three response format conditions separately and found that in every condition people were less confident in the 01 conflict category than in the baseline, $b < -21.4$, $t < -9.7$, $p < 0.0001$. With respect to reaction times, we found that conflict improved model fit significantly, χ^2 (1) = 28.6, $p < 0.0001$, as well as the main effect of condition, χ^2 (3) = 566.4, $p < 0.0001$, but not their interaction, χ^2 (5) = 2.1, $p = 0.34$. It took participants longer to give a response in the 01 conflict category than in the baseline, $b = 0.09$, t (324.5) = 6.8, $p < 0.0001$.

Taken together, the conflict detection analysis on the confidence and latency data indicates that by and large participants showed decreased response confidence and increased response times (in contrast with the no-conflict baseline) after having given their first, intuitive response on the conflict problems in all direction of change categories. This supports the hypothesis that participants were always being faced with two conflicting intuitive responses when solving the conflict bat-and-ball problems. In other words, results imply that 11 responders also activate a heuristic “10 cents” intuition in addition to the logical correct “5 cents” response they selected. Likewise, 00 responders also seem to detect that there is an alternative to the incorrect “10 cents” response. Although this points to some minimal error sensitivity (De Neys et al., 2013; Johnson et al., 2016) among incorrect responders, it does not imply that incorrect responders also realize that the correct response is “5 cents” (Travers et al., 2016). The error sensitivity or increased doubt for incorrect responders might result from a less specific intuition (e.g., maybe incorrect responders doubted that their “10 cents” was correct without knowing that the correct response was “5 cents”). More generally speaking, it is possible that the correct intuition differs in strength and/or specificity for correct and incorrect responders. Clearly, the present study was designed and optimized to draw conclusions about the nature of correct

responders' intuitions. Claims with respect to the nature of incorrect responders' intuitions remain speculative and will need further validation in future studies.

Finally, visual inspection also clearly shows that the conflict effects were much larger for the 10 and 01 cases than for the 11 and 00 ones. A contrast analysis[11] that tested this trend directly indicated that this trend was significant for confidence data, $Z = -15.4$, $p < 0.0001$, ($r = 0.11$ for the no-change group, while $r = 0.5$ for the change group), and reaction times, $Z = -2.35$, p (one-tailed) $= 0.009$, ($r = 0.07$ for no-change and $r = 0.13$ for change group). This pattern suggests that although reasoners might be generating two intuitive responses and are being affected by conflict between them in all cases, this conflict is much more pronounced in cases where people subsequently change their answer. This tentatively suggests that it is this more pronounced conflict experience that makes them subsequently change their answer (Bago & De Neys, 2017; Thompson et al., 2012).

[11] For this contrast analysis, we first calculated the r effect sizes out of t-values (Rosnow & Rosenthal, 2003). As a next step we used Fisher r-to-z transformation to assess the statistical difference between the two independent r-values. We used the following calculator for the z-transformation and p-value calculation: http://vassarstats.net/rdiff.html

Table S1. Average confidence differences (SD) at the initial response stage between the baseline (11 responses on no-conflict problems) and conflict problems for each direction of change category.

	Response format	11	00	10	01
Study1	2 response	8.4 (3.4)	2.5 (1.9)	52 (9.97)	25.7 (7.6)
Study2a	2 response	14.7 (5.6)	1.7 (1.2)	50.1 (8.1)	19.4 (8.7)
Study2b	4 response	9.6 (4.9)	4.8 (1.5)	42.4 (48.5)	41.8 (10.6)
Study3a	2 response	27.8 (7.7)	7.9 (3.5)	65.7 (7.8)	29.1 (7.6)
Study3b	4 response	12.5 (5.4)	9.8 (3.4)	-	32.6 (10.2)
Study4	Free response	10.3 (6.2)	1.2 (0.9)	-	34.3 (32.5)
Study5	Free response	8.6 (3.5)	9.2 (3.6)	95.7 (-)	55 (10.2)
Average	2 response	15.7 (2.8)	2.4 (1)	55.0 (5.1)	25.7 (4.5)
	4 response	10.8 (3.7)	5.7 (1.4)	42.3 (48.5)	36.9 (7.3)
	Free response	9.8 (3)	3.7 (1.6)	97.4 (-)	52.7 (9.8)
Overall average		12.3 (1.9)	3.8 (0.7)	56.1 (5.1)	33.5 (3.7)

Table S2. Average reaction time differences in ms (SD) at the initial response stage between the baseline (11 responses on no-conflict problems) and conflict problems for each direction of change category. Note that averages are based on geometrical means.

	Response format	11	00	10	01
Study1	2 response	-300 (196.4)	-200 (124.8)	-350 (521.5)	-770 (290.7)
Study2a	2 response	-230 (291.6)	-40 (90.3)	180 (388.7)	-60 (373.8)
Study2b	4 response	-550 (239)	-170 (93.6)	200 (1944)	-230 (472.3)
Study3a	2 response	-510 (302.4)	-60 (191.6)	-470 (473)	-390 (334.7)
Study3b	4 response	-470 (276.9)	50 (199.3)	-	-900 (386.9)
Study4	Free response	-230 (389.3)	-220 (183.8)	-	-3230 (658.1)
Study5	Free response	-300 (235.4)	-210 (187.2)	-2340 (-)	-1640 (326)
Average	2 response	-400 (143.5)	-60 (68.7)	-120 (267.4)	-490 (195.1)
	4 response	-580 (181.1)	-120 (84.8)	270 (1943.8)	-670 (307.7)
	Free response	-440 (202.5)	-100 (131.2)	-2750 (-)	-2170 (291.6)
Overall average		-720 (113.8)	-30 (55.7)	270 (269.2)	-760 (175.4)

C. Data for no-conflict control problems

The tables in this section give an overview of the direction of change (Table S3), stability (Table S4), and justification data (Table S5 & S6) on the no-conflict control problems.

Table S3. Frequency of direction of change categories for no-conflict control problems in Study 1-5. The raw number of trials in each category is presented between brackets.

	Response format	11	00	10	01
Study1	2 response	93.0% (281)	3.6% (11)	0.3% (1)	3.0% (9)
Study2a	2 response	93.9% (504)	2.8% (15)	0.7% (4)	2.6% (14)
Study2b	4 response	93.3% (457)	3.3% (16)	1.6% (8)	1.8% (9)
Study3a	2 response	94.9% (129)	1.5% (2)	1.5% (2)	2.2% (3)
Study3b	4 response	94.1% (128)	1.5% (2)	1.5% (2)	2.9% (4)
Study4	Free response	96.4% (132)	0.77% (1)	-	2.9% (4)
Study5	Free response	92.5% (149)	0.6% (1)	1.2% (2)	5.6% (9)
Average	2 response	93.7% (914)	2.9% (28)	0.7% (7)	2.7% (26)
	4 response	93.5% (585)	2.9% (18)	1.6% (10)	2.1% (13)
	Free response	94.5% (281)	0.7% (2)	0.7% (2)	4.4% (13)
Overall average		93.8% (1780)	2.5% (48)	1.0% (19)	2.7% (52)

Table S4. Frequency of stability index values on no-conflict control problems in Study 1-5. The raw number of participants in each category is presented between brackets.

	Response format	<33%	50%	66%	75%	100%	Average stability
Study1	2 response	1.2% (1)	2.3% (2)	4.6% (4)	5.8% (5)	86.2% (75)	95.1%
Study2a	2 response	-	1.3% (2)	5.0% (8)	5.7% (9)	88.1% (140)	96.3%
Study2b	4 response	-	2.8% (4)	4.9% (7)	4.2% (6)	88.2% (127)	96.0%
Study3a	2 response	2.4% (1)	4.9% (2)	4.9% (2)	2.4% (1)	85.4% (35)	94.7%
Study3b	4 response	-	4.9% (2)	4.9% (2)	4.9% (2)	85.4% (35)	96.4%
Study4	Free response	-	-	5.1% (2)	7.7% (3)	87.2% (34)	93.0%
Study5	Free response	-	2.3% (1)	2.3% (1)	20.5% (9)	75.0% (33)	95.0%
Average	2 response	0.7% (2)	2.1% (6)	4.9% (14)	5.2% (15)	87.2% (251)	95.5%
	4 response	-	3.2% (6)	4.9% (9)	4.3% (8)	87.6% (162)	95.6%
	Free response	-	1.2% (1)	3.6% (3)	14.5% (12)	80.7% (67)	94.6%
Overall average		0.4% (2)	2.3% (13)	4.7% (26)	6.3% (35)	86.3% (480)	95.4%

Table S5. Frequency of different justification categories for no-conflict control problems in Study 6. The raw number of justifications in each category is presented between brackets.

Justification	Initial response		Final response	
	Correct	Incorrect	Correct	Incorrect
Correct math	39.6% (19)	-	56.5% (48)	-
Incorrect math	-	-	-	20% (1)
Unspecified math	10.4% (5)	-	23.5% (20)	-
Hunch	6.3% (3)	-	1.2% (1)	-
Guess	6.3% (3)	25% (3)	1.2% (1)	40% (2)
Previous	6.3% (3)	25% (3)	3.5% (3)	-
Other	31.3% (15)	50% (6)	14.1% (12)	40% (2)

Table S6. Frequency of different justification categories for no-conflict control problems in Study 7. The raw number of justifications in each category is presented between brackets.

Response format	Justification	Initial response		Final response	
		Correct	Incorrect	Correct	Incorrect
2 response	Correct math	51.9% (41)	-	82.8% (72)	-
	Incorrect math	-	-	-	-
	Unspecified math	17.7% (12)	12.5% (1)	12.6% (11)	-
	Hunch	19% (15)	25% (2)	3.4% (3)	-
	Guess	10.2% (8)	37.5% (3)	-	-
	Other	1.2% (1)	25% (2)	1.1% (1)	-
Free response	Correct math	68.7% (68)	12.5% (1)	95.3% (102)	-
	Incorrect math	-	-	-	-
	Unspecified math	2% (2)	-	0.9% (1)	-
	Hunch	15.2% (15)	12.5% (1)	0.9% (1)	-
	Guess	14.1% (14)	62.5% (5)	1.9% (2)	-
	Other	-	12.5% (1)	0.9% (1)	-

D. Full procedure reading pre-tests Study 1 and Study 2

Reading pre-test Study 1. Before we ran the main study we also recruited an independent sample of 64 participants for a reading pre-test (31 female, Mean age = 38.9 years, SD = 13.1 years). Participants were recruited via the Crowdflower platform, and they received $0.10. A total of 41% of the subjects reported high school as highest completed educational level, and 58% reported having a post-secondary education degree (1% reported less than high school). The basic goal of the reading pre-test was to determine the response deadline which could be applied in the main reasoning study. The idea was to base the response deadline on the average reading time in the reading test. Note that dual process theories are highly underspecified in many aspects (Kruglanski, 2013); they argue that System 1 is faster than System 2, but do not further specify how fast System 1 is exactly (e.g., System 1 < x seconds). Hence, the theory gives us nu unequivocal criterion on which we can base our deadline. Our "average reading time" criterion provides a practical solution to define the response deadline. The rationale here was very simple; if people are allotted the time they need to simply read the problem, we can be reasonably sure that System 2 engagement is minimal. Thus, in the reading pre-test, participants were presented with the same items as in the reasoning study. They were instructed to read the problems and randomly click on one of the answer options. Of course, we wanted to avoid that participants would spontaneously engage in any type of reasoning in the pre-test. Therefore, the answer options were randomly selected numbers (to which we drew participant's attention in the instructions) to make it less likely that reading participants would try to solve the problems. The general instructions were as follows:

> **Welcome to the experiment!**
> **Please read these instructions carefully!**
> This experiment is composed of 8 questions and 1 practice question. It will take 3 minutes to complete and it demands your full attention. You can only do this experiment once. In this task we'll present you with a set of problems we are planning to use in future studies. Your task in the current study is pretty simple: you just need to read these problems. We want to know how long people need on average to read the material. In each problem you will be presented with two answer alternatives. You don't need to try to solve the problems or start thinking about them. Just read the problem and the answer alternatives and when you are finished reading you randomly click on one of the answers to advance to the next problem. In each problem you will be presented with two answer alternatives. These answer alternatives are simply randomly generated numbers. You don't need to try to solve the problems or start thinking about them. The only thing we ask of you is that you stay focused and read the problems in the way you typically would. Since we want to get an accurate reading time estimate please avoid whipping your nose, taking a phone call, sipping from your coffee, etc. before you finished reading. At the end of the study we will present you with some easy verification questions to check whether you actually read the problems. This is simply to make sure that participants are complying with the instructions and actually read the problems (instead of clicking through them without paying attention). No worries, when you simply read the problems, you will have no trouble at all at answering the verification questions.
>
> You will receive $0.10 for completing this experiment.
> Please confirm below that you read these instructions carefully and then press the "Next" button.

To make sure that participants would actually read the problems, we informed subjects that they would be asked to answer two – very easy - verification questions at the end of the

experiment to check whether they read the material. The verification questions could be easily answered even by a very rough reading. The following illustrates the verification question:

> We asked you to read a number of problems.
> Which one of the following pair of goods were NOT presented during the task?
- o A laptop and a mouse
- o A pencil and an eraser
- o An apple and an orange
- o A banana and a magazine

The correct answers were clearly different from the goods which were presented during the task. A total of 84.4% of the participants solved both verification questions correctly, and only the data from these participants was analysed.

As in the main experiment, items were presented serially. First, the first sentence of the problem was presented for 2000 ms. Next, the full problem appeared on the screen. Reading times were measured from the presentation of the full problem. The average reading time of the sample was M = 3.87 sec, SD = 2.18 sec. Note that raw reaction time data were first logarithmically transformed prior to analysis. Mean and standard deviation were calculated on the transformed data, and then they were back-transformed into seconds. We wanted to give the participants some minimal leeway[12], thus we rounded the average reading time to the closest higher natural number; the response deadline was therefore set to 4 seconds.

Reading pre-test Study 2. Half of the participants were presented with four response options in Study 2. Since reading through more options will in itself take more time, we decided to run a new reading pre-test with the 4-option format. To this end an additional 23 participants were recruited (16 female, mean age = 40.8 years, SD = 15.2 years) via Crowdflower. They received $0.10 for participation. A total of 48% of the participants reported high school as highest completed educational level, and 52% reported having a post-secondary education degree. As in Study 1, the four response options in the pre-test were randomly generated numbers. Except for the number of response options the pre-test was completely similar to the 2-option Study 1 pre-test. Participants were also presented with the same verification questions, which were correctly solved by 74.9% of the participants. Only data from participants who responded correctly to both verification questions was analysed. Prior to reaction time analysis, raw reaction times were log-transformed. Mean and standard deviations were calculated on the log-transformed data, and they were back-transformed after calculation. The mean reading time in the pre-test sample was 4.3 s (SD = 2 s). As in Study 1, we rounded the deadline to nearest higher natural number. Hence, the time limit in the 4-option format was set to 5 s (vs 4 s in the 2-option format).

[12] This also helped to account for minor language differences since participants in the main study would solve Hungarian translations of the English problems.

E. Additional analyses and data

Table S7. Frequency of direction of change categories in study 1-5 for the first conflict problem that participants solved only. Raw number of trials are in brackets.

Study	Response format	Direction of change category				Non-correction (11/11+01)
		11	00	10	01	
Study1	2 response	5.7% (5)	71.6% (63)	6.8% (6)	15.9% (14)	26%
Study2a	2 response	4.2% (7)	89.2% (148)	2.4% (4)	4.2% (7)	50%
Study2b	4 response	8.1% (12)	89.2% (132)	-	2.7% (4)	75%
Study3a	2 response	2.4% (1)	69.1% (29)	9.5% (4)	19.1% (8)	11%
Study3b	4 response	9.3% (4)	76.7% (33)	-	14% (6)	40%
Study4	Free response	10% (4)	85% (34)	-	5% (2)	66.7%
Study5	Free response	9.1% (4)	68.2% (30)	-	22.7% (10)	29%
Average	2 response	4.4% (13)	81.1% (240)	4.7% (14)	9.8% (29)	30.1%
	4 response	8.4% (16)	86.4% (165)	-	5.2% (10)	61.5%
	Free response	9.5% (8)	76.2% (64)	-	14.3% (12)	40%
Overall average		6.5% (37)	82.1% (469)	2.5% (14)	8.9% (51)	42%

Table S8. Frequency of each direction of change category for the first conflict item only in the justification studies (Study 6-7). Raw number of trials are in brackets.

Study	Response format	Direction of change category				Non-correction (11/11+01)
		11	00	10	01	
Study 6	2 response	8.2% (4)	53.1% (26)	16.3% (8)	22.5% (11)	26.7%
Study 7a	2 response	18.8% (9)	58.3% (28)	10.4% (5)	12.5% (6)	60%
Study 7b	Free response	12.5% (7)	44.6% (25)	1.8% (1)	41.1% (23)	23.3%

Table S9. Frequency of conflict problem justifications for different direction of change categories in Study 6 and 7 (raw number of justifications in brackets).

Study	Justification	Initial response				Final response			
		11	00	10	01	11	00	10	01
Study 6 2 response	Correct math	30% (6)	-	-	7.1% (1)	55% (11)	6.7% (3)	-	69.2% (9)
	Incorrect math	-	25.5% (12)	-	21.4% (3)	-	37.8% (17)	11.1% (1)	-
	Unspecified math	-	19.1% (9)	-	-	5% (1)	37.8% (17)	77.8% (7)	7.7% (1)
	Hunch	5% (1)	4.3% (2)	-	14.3% (2)	5% (1)	-	-	7.7% (1)
	Guess	20% (4)	10.6% (5)	66.7% (6)	7.1% (1)	5% (1)	4.4% (2)	-	-
	Previous	30% (6)	6.4% (3)	-	7.1% (1)	10% (2)	2.2% (1)	-	-
	Other	15% (3)	34% (16)	33.3% (3)	42.9% (6)	20% (4)	11.1% (5)	11.1% (1)	15.4% (2)
Study 7 2 response	Correct math	12.5% (2)	2.3% (1)	-	-	93.8% (15)	2.3% (1)	-	100% (10)
	Incorrect math	-	40.1% (18)	-	20% (2)	-	65.9% (29)	83.3% (5)	-
	Unspecified math	12.5% (2)	11.4% (5)	-	-	6.3% (1)	18.2% (8)	-	-
	Hunch	56.3% (9)	22.7% (10)	16.7% (1)	50% (5)	-	2.3% (1)	-	-
	Guess	18.8% (3)	22.7% (10)	83.3% (5)	30% (3)	-	11.4% (5)	-	-
	Other	-	-	-	-	-	-	16.7% (1)	-
Study 7 Free response	Correct math	30.8% (4)	-	-	-	84.6% (11)	2.3% (1)	-	93.3% (28)
	Incorrect math	-	46.5% 20	-	26.7% (8)	-	69.8% (30)	50% (1)	-
	Unspecified math	7.7% (1)	7% (3)	-	-	7.7% (1)	16.3% (7)	50% (1)	3.3% (1)
	Hunch	38.5% (5)	23.3% (10)	100% (2)	40% (12)	-	4.7% (2)	-	-
	Guess	15.4% (2)	21% (9)	-	26.7% (8)	-	7% (3)	-	-
	Other	7.7% (1)	2.3% (1)	-	6.7% (2)	7.7% (1)	-	-	3.3% (1)

Table S10. Frequency of "11" stability on conflict items in Study 1-5. The raw number of participants for each value is presented between brackets.

Study	Response format	Stability index value					Average stability
		<33%	50%	66%	75%	100%	
Study1	2 response	18.2% (4)	18.2% (4)	22.7% (5)	22.7% (5)	18.2% (4)	65.2%
Study2a	2 response	23.1% (3)	30.8% (4)	7.7% (1)	-	38.5% (5)	66%
Study2b	4 response	38.1% (8)	9.5% (2)	9.5% (2)	4.8% (1)	38.1% (8)	64.3%
Study3a	2 response	18.2% (2)	36.4% (4)	9.1% (1)	36.4% (4)	-	57.6%
Study3b	4 response	16.7% (2)	8.3% (1)	33.3% (4)	8.3% (1)	33.3% (4)	70.8%
Study4	Free response	-	16.7% (1)	33.3% (2)	-	50% (3)	80.6%
Study5	Free response	5.6% (1)	33.3% (6)	16.7% (3)	22.2% (4)	22.2% (4)	68.1%
Average	2 response	19.6% (9)	26.1% (12)	15.2% (7)	19.6% (9)	19.6% (9)	63.6%
	4 response	30.3% (10)	9.1% (3)	18.2% (6)	6.1% (2)	36.4% (12)	66.7%
	Free response	4.2% (1)	29.2% (7)	20.8% (5)	16.7% (4)	29.2% (7)	71.2%
Overall average		19.4% (20)	21.4% (22)	17.5% (18)	14.6% (15)	27.2% (28)	66.3%

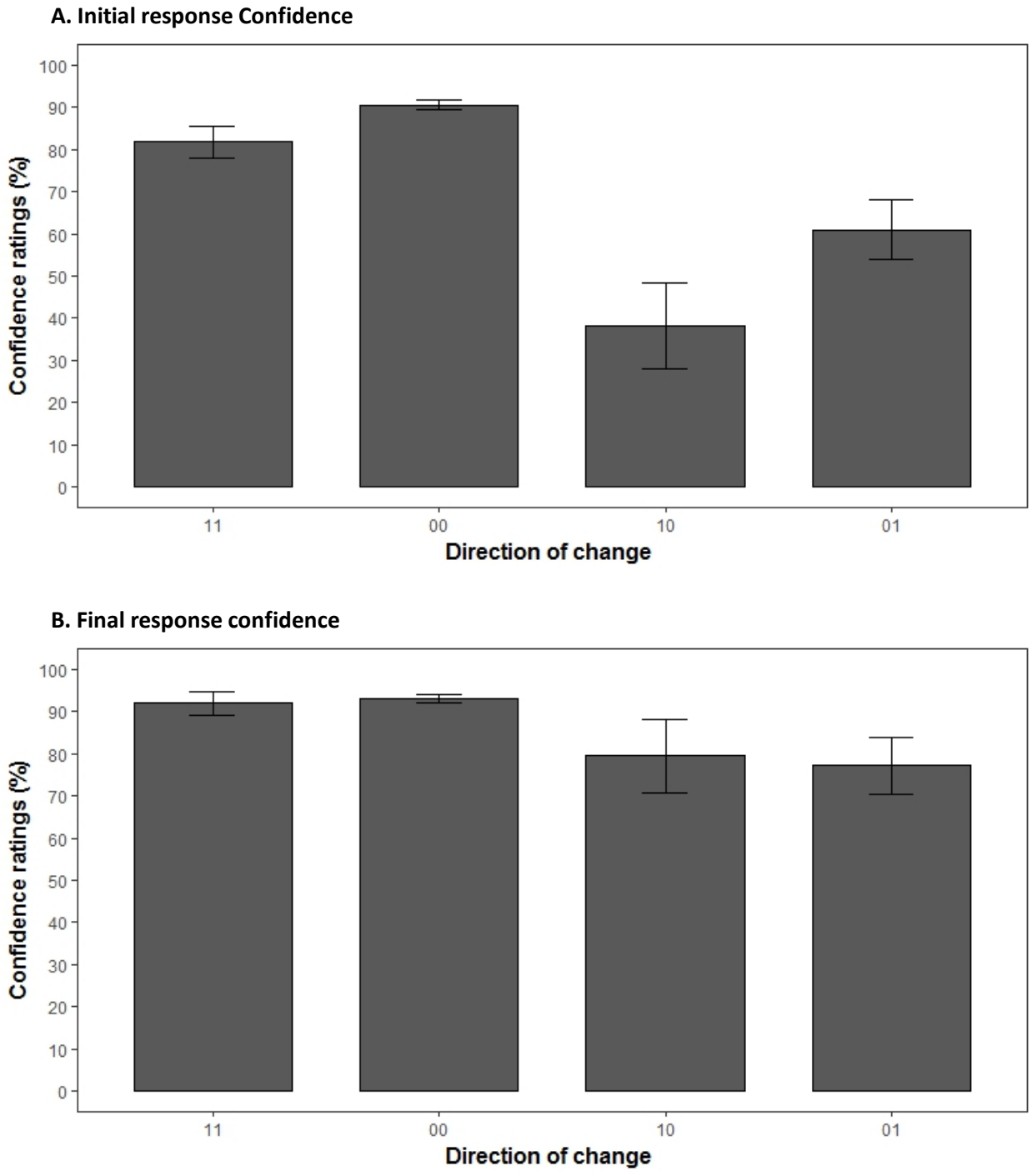

Figure S1. Mean initial (A.) and final (B.) conflict problem response confidence ratings averaged across Study 1-5. Error bars are 95% confidence intervals.