Second-guess: Testing the specificity of error detection in the bat-and-ball problem

Bence Bago, Matthieu Raoelison, Wim de Neys

SECOND-GUESS: TESTING THE SPECIFICITY OF ERROR DETECTION IN THE BAT-AND-BALL PROBLEM

Bence Bago*, Matthieu Raoelison, & Wim De Neys

LaPsyDE (CNRS Unit 8240), Sorbonne - Paris Descartes University, Paris, France

*Corresponding author: Bence Bago
LaPsyDÉ (Unité CNRS 8240, Université Paris Descartes)
Sorbonne - Labo A. Binet
46, rue Saint Jacques
75005 Paris
France

bencebagok@gmail.com

ABSTRACT

In the last decade conflict detection studies in the reasoning and decision-making field have suggested that biased reasoners who give an intuitive response that conflicts with logico-mathematical principles can often detect that their answer is questionable. In the present studies we introduced a second guess paradigm to test the nature and specificity of this error or conflict signal. Participants solved the bat-and-ball problem and were allowed to make a second guess after they had entered their answer. Three studies in which we used a range of second guess elicitation methods show that biased reasoners predominantly give second guesses that are smaller than the intuitively cued heuristic response ("10 cents"). Findings indicate that although biased reasoners do not know the exact correct answer ("5 cents") they do correctly grasp that the right answer must be smaller than the intuitively cued "10 cents" answer. This suggests that reasoners might be savvier about their errors than traditionally assumed. Implications for the conflict detection and dual process literature are discussed.

Keywords: Reasoning; Heuristics and biases; Conflict detection; Dual process theory; Bat-and-ball problem

INTRODUCTION

Studies on reasoning and decision making have long established that human thinking is often biased. From our answers to logical and probabilistic reasoning tasks to juror's death penalty judgments and our stock-market picks, people seem to base their judgments on intuitive rules-of-thumb instead of on more demanding, deliberate thinking (e.g., Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006; Kahneman, 2011; Oster & Koesterich, 2013). Although this intuitive, so-called heuristic thinking can be useful it will sometimes cue responses that conflict with logical, probabilistic, mathematical or other normative principles. In these cases, intuitions can bias our inferencing. Consider, for example, the infamous bat-and-ball problem (Frederick, 2005).

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

Intuitively, the answer "10 cents" immediately springs to mind because we naturally tend to parse the $1.10 in $1 and 10 cents (Kahneman, 2011). Indeed, "10 cents" is the answer that the vast majority of even highly educated university students come up with (Bourgeois-Gironde & Van Der Henst, 2009). However, although it is intuitively appealing the answer is not correct. 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. 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.

The biased responding on the bat-and-ball problem is quite striking. In theory, solving the bat-and-ball problem shouldn't be too hard. It boils down to solving the basic algebraic equation "X + Y = 1.10, Y = 1 + X, Solve for X" - something all educated adults have done at length in their high school math classes (Hoover & Healy, 2017). Nevertheless, the intuitive appeal of the "10 cents" answer seems to have an irresistible pull on people's thinking and leads them astray.

To pinpoint the nature of intuitive bias numerous studies in the last decade have focused on the role of conflict detection during reasoning (e.g., De Neys & Glumicic, 2008; Pennycook, Fugelsang, & Koehler, 2015; Thompson & Johnson, 2014). These studies test whether biased reasoners who give an intuitive response that conflicts with logico-mathematical principles, show any sensitivity to this conflict. In other words, are biased reasoners blind heuristic thinkers who are completely insensitive to the fact that their response is logically questionable or do they show some minimal error sensitivity? The answer to this question has far-stretching implications for our view of human rationality and our theories about the interaction of intuition and deliberation in thinking (De Neys, 2012, 2017; Evans, 2007; Pennycook et al., 2015).

To address the question empirically, conflict or bias detection studies typically present participants with conflict and control "no-conflict" versions of classic reasoning tasks. In the traditional conflict versions – such as the above bat-and-ball problem - the intuitively cued heuristic response conflicts with the correct response. In the control, no-conflict versions the intuitively cued heuristic response is made coherent with the correct logico-mathematical response. For example, a no-conflict control problem of the bat-and-ball problem might read:

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

Obviously, in this control version the intuitive splitting of $1.10 and selection of the "10 cents" answer is also mathematically correct.

To test reasoners' conflict or error detection sensitivity, studies contrast participants' processing of the conflict and control versions. The key difference between the conflict and control problems is the fact that the intuitively cued heuristic response happens to be incorrect on the conflict problem. If biased reasoners are sensitive to the erroneous nature of their answer, one can expect that this detection will affect their processing (e.g., Botvinick, 2007). Although there have been a number of negative findings (e.g., Ferreira, Mata, Donkin, Sherman, & Ihmels, 2016; Mata, Ferreira, Voss, & Kollei, 2017; Pennycook, Fugelsang, & Koehler, 2012) such processing effects have been observed in the bulk of these studies. In a wide range of

tasks biased reasoners who solve conflict versions typically need more time to make a decision (e.g., Bonner & Newell, 2010; De Neys & Glumicic, 2008; Pennycook, Trippas, Handley, & Thompson, 2014; Stupple, Ball, Evans, & Kamal-Smith, 2011; Villejoubert, 2009), are less confident about the correctness of their response (e.g., Bago & De Neys, 2017; De Neys, Cromheeke, & Osman, 2011; Gangemi, Bourgeois-Gironde, & Mancini, 2015; Johnson, Tubau, & De Neys, 2016; Thompson & Johnson, 2014), and show increased activation of brain areas assumed to mediate conflict and error monitoring (e.g., Anterior Cingulate Cortex; e.g., De Neys, Vartanian, & Goel, 2008; Simon, Lubin, Houdé, & De Neys, 2015; Vartanian et al., 2018) compared to when they solve control versions.

In sum, the empirical conflict detection studies present considerable support for the conclusion that even biased, incorrect responders to traditional (i.e., conflict) reasoning problems often detect that the intuitively cued heuristic answer is questionable. When people give a heuristic response that conflicts with logico-mathematical principles, they seem to be picking up on this conflict at some level. However, the problem lies with the "at some level" qualification in the previous sentence. The empirical studies indicate that there is evidence for the presence of a conflict or error signal. However, the precise nature of the signal is not clear (e.g., Aczel, Szollosi, & Bago, 2016; Handley & Trippas, 2015; Johnson et al., 2016; Koriat, 2017; Singmann, Klauer, & Kellen, 2014; Stupple, Pitchford, Ball, Hunt, & Steel, 2017; Szollosi, Bago, Szaszi, & Aczel, 2017; Travers, Rolison, & Feeney, 2016; Villejoubert, 2009). Consider, for example, the bat-and-ball problem. Conflict detection studies indicated that biased reasoners doubt their "10 cents" response in case it conflicts with the correct response (e.g., De Neys, Rossi, & Houdé, 2013; Gangemi et al., 2015; Johnson et al., 2016; Szollosi et al., 2017). But where does this doubt come from? One possibility is that people have a highly specific error or conflict signal. Reasoners might have computed both the "10 cents" and "5 cents" response. Hence, they experience a conflict between the incorrect and correct response. Biased reasoners would be in doubt between the two options but find the heuristic "10 cents" answer relatively more compelling. Critically, although people still end up being biased, they would know that the possible alternative answer is "5 cents" in this case. Alternatively, the conflict signal might be non-specific. That is, people might detect that the "10 cents" is questionable without having

any further clue about what the correct response is. For example, people might know that the "10 cents" response might be incorrect because they realize they did not process the problem premises properly (Johnson et al., 2016; Szollosi et al., 2017). However, they might not have any further insight about what the correct response is.

This problem was perhaps most clearly illustrated by Travers et al. (2016). In their study Travers et al. adopted a mouse-tracking paradigm. Different response options were presented in each of the corners of the screen (e.g., "10 cents", "5 cents") and participants had to move the mouse pointer from the center of the screen towards the response option of their choice to indicate their decision. In the mouse-tracking paradigm researchers typically examine the curvature in the mouse movement to test whether the non-chosen response exerts some competitive "pull" or attraction over the chosen response (Spivey, Grosjean, & Knoblich, 2005). One can use this attraction as a measure of conflict detection. That is, if incorrect responders are considering the correct response, they should tend to slightly move towards it resulting in a more curved mouse trajectory[1]. Although Travers et al. found that correct responders showed attraction to the incorrect "10 cents" response, incorrect responders did not show attraction to the correct "5 cents" response. As Travers et al highlighted, this might imply that contrary to other conflict detection findings with the bat-and-ball problem, biased responders do not show any error sensitivity. Or, it might simply indicate that biased reasoners' conflict detection is non-specific in nature. If reasoners detect that the "10 cents" response is questionable but do not know that "5 cents" is correct, it would not be surprising that they show no specific attraction towards the "5 cents" option.

The inconclusiveness in the Travers et al. (2016) study illustrates why both proponents and opponents of the idea that reasoners are conflict sensitive have stressed that it is crucial for further theory development to pinpoint the specific nature of the process (Aczel et al., 2016; Bago & De Neys, 2017, 2018; Ferreira et al., 2016; Johnson et al., 2016; Newman, Gibb, & Thompson, 2017; Singmann et al., 2014). In the present study we start to address this issue.

[1] For the record, note that Travers et al. (2016) actually used a slightly different analysis procedure. Instead of looking at the curvature of a single mouse movement, they tracked the position of the cursor over a long window of time (up to 60 s). Conclusions are conceptually similar.

We therefore introduce a second guess paradigm. First, we present participants with the bat-and-ball problem and ask them to generate a response. Next, we ask them to make a second guess and ask them to choose among different options, say, “1 cent”, “5 cent”, or “15 cent”. Critically, the intuitively cued response is not among the second guess options. People’s second guess choice should reflect the specificity of their conflict detection. Imagine that biased “10 cents” reasoners have a highly specific conflict signal. They have computed both the “10 cents” and “5 cents” responses but were in doubt and initially found the “10 cents” more compelling. Now, when being asked for a second guess, given that the “10 cents” option is no longer available, they should clearly opt for the “5 cents” response. Alternatively, imagine that reasoners’ conflict signal is non-specific in nature: they know there might be an alternative to the “10 cents” response, but they have no clue about what this alternative is. Consequently, when second guessing they will need to - literally - guess, and the different second guess options should be chosen with equal frequency. Interestingly, in addition to a non-specific and high-specific error signal, one might envisage other “intermediate” levels of error detection specificity. Imagine that biased reasoners did not manage to compute the correct “5 cents” response but they did detect that the “10 cents” is questionable because it implies a total cost of $1.20 which is too high. That is, people might understand that “10 cents + ($1 + 10 cents) > $1.10” and realize that therefore the correct response must be smaller than “10 cents”. We might call this a medium level conflict signal. Clearly, in this case people might not know that the correct answer is precisely “5 cents” but they would at least know that the correct answer can only be found “below” the intuitively appealing “10 cents”. Consequently, when second guessing they should manage to avoid the higher “foil” option (e.g., “15 cents”) but be further indifferent between the “1 cent” and “5 cents” options since both satisfy their constraint. This illustrates how we can gain insight into the specificity of the error or conflict signal by examining the distribution of second guess responses. We present three studies that adopted this approach.

STUDY 1

Method

Participants

In Study 1, 231 Hungarian undergraduate students (176 female, Mean age = 22.9 years, SD = 3.4 years) from the Eotvos Lorand University of Budapest were tested. Participants received course credit for taking part.

Material

Reasoning items. Participants were presented with one standard conflict and one control no-conflict version of the bat-and-ball problem. Participants completed the study online. As in previous studies (e.g., De Neys et al., 2013; Johnson et al., 2016), we modified the superficial item content of the two problems (i.e., one problem stated that a pencil and eraser together cost $3.30, the other that a magazine and banana together cost $4.40). To make sure that the specific item content did not affect the findings, the item content and conflict status of the problems were completely counterbalanced. Presentation order of the control and no-conflict problems was also randomized. Participants typed their answer in a blank box with the label "cent(s)" next to it that appeared on screen under the problem. Here are the full English translations of the conflict and no-conflict problems that we adopted:

> Conflict versions:
> A pencil and eraser together cost $3.30. The pencil costs $3 more than the eraser. How much does the eraser cost? *(correct response = 15 cents, heuristic response = 30 cents)*
>
> A magazine and banana together cost $4.40. The magazine costs $4 more than the banana. How much does the banana cost? *(correct response = 20 cents, heuristic response = 40 cents)*
>
> No-conflict versions:
> A pencil and eraser together cost $3.30. The pencil costs $3. How much does the eraser cost? *(correct/heuristic response = 30 cents)*
>
> A magazine and banana together cost $4.40. The magazine costs $4. How much does the banana cost? *(correct/heuristic response = 40 cents)*

Immediately after participants entered their answer the problem disappeared from the screen and participants were asked to indicate how confident they were that their response was correct by typing a number between 0 (totally unsure) and 100 (completely certain) in a

blank box. As in previous studies we recorded the confidence in the conflict and no-conflict answers (in addition to the problem decision time) to measure participants' conflict detection sensitivity (e.g., Johnson et al., 2016)[2].

To familiarize participants with the response box format they first saw a simple and unrelated math story practice problem where they had to enter a numerical response and confidence estimate in the response box.

Second guess question. After participants had entered their answer and confidence estimate they were presented with the second guess question. The question stated: "Imagine that your answer to the problem you just solved doesn't turn out to be right. Which one of the following options would you pick as your second guess?". Participants were presented with four numerical options which were shown listed beneath each other (in a randomly determined order for each participant). This is what the second guess screen looked like:

> Imagine that your answer to the problem you just solved doesn't turn out to be right. Which one of the following options would you pick as your second guess?
>
> o 10 cents
> o 20 cents
> o 60 cents
> o 70 cents

The second guess options on the conflict problems were conditional on the correctness of the first answer participants had entered. We first explain the critical case of a biased participant who failed to give the correct response to a conflict problem. The four second guess response options were constructed as follows: One option was the correct response, a second option was a lower foil (i.e., correct response – 10), a third option was a higher foil (i.e., heuristic + correct response), and a fourth option was an extremer high foil (i.e., higher foil + 10). For the problem version that totaled to $4.40 (e.g., correct response = 20 cents, heuristic response = 40 cents) this resulted in the following second guess options: 10 cents (lower foil), 20 cents (correct

[2] We also recorded the confidence measure response time (i.e., the time needed to make a confidence judgment) but did not analyze these data given recent findings that question the robustness of this measure as a conflict detection index (Frey, Johnson, & De Neys, 2018).

response), 60 cents (higher foil), 70 cents (extreme foil). For the problem version that totaled to $3.30 (i.e., correct response = 15 cents, heuristic response = 30 cents) the second guess options were: 5 cents (lower foil), 15 cents (correct), 45 cents (higher foil), 55 cents (extreme foil).

We should stress that special care was taken in the selection of these specific response alternatives. The second guess questioning can only inform us about the specificity of the conflict signal if confounding response biases are excluded. Our main concern was to avoid simple priming or oddball effects. The following a priori rules were established: 1) all response options end in 5 ($3.30 problem version) or 0 ($4.40 problem version), 2) the heuristic response option is the numerical mean of the four options, and 3) the low foil and correct option vs high foil and extreme high foil lie at symmetrical (opposite) numeral distance from the heuristic response (e.g., with heuristic response a, the four options must be a-x/a-y/a+x/a+y).

It will be clear that our second guess manipulation was designed and optimized to draw conclusions about incorrect conflict responders' inferencing. For participants who solved a conflict problem correctly the correct second guess response option was replaced with the heuristic response. On no-conflict control problems (on which the heuristic response is also correct), we always presented the same four options as for an incorrectly solved conflict version. Hence, the a priori second guess option selection rules did not apply here.

For convenience, in our results and discussion sections we will always illustrate the second guess options by referring to numerical values based on the original bat-and-ball problem (e.g., "1 cent" = low foil, "5 cents" = correct option, "10 cents" = heuristic option, etc.). Given the familiarity of the bat-and-ball problem we believe this keeps the exposition maximally accessible. However, the reader should bear in mind that these values are used for illustration only.

One might wonder why we went through the trouble of optimizing the presented second guess options and did not simply ask participants to generate a second guess themselves (i.e., free-response format). Note that such an open-ended, free-response second guess format will be adopted in Study 3. But we initially opted against this procedure because it is well established that people have a strong tendency to stick to their initial response. Various experimental paradigms indicate that people are reluctant to change their initial answer when

given the chance (e.g., De Neys & Verschueren, 2006; Gilovich, Medvec, & Chen, 1995; Thompson, Prowse Turner, & Pennycook, 2011). With an open response format people can be tempted to enter the same response. Such a bias would render the second guess uninformative. Our forced choice approach in which the initially selected response is not among the response options sidesteps this potential complication (see also the structured elicitation in Study 2 for an alternative approach and the free-response format findings of Study 3).

Participants clicked on the option of their choice to select a second guess response. Afterwards they were allowed to take a short break and continued with the subsequent problem.

Exclusion criterion. At the end of the experiment participants answered standard demographic questions and were also presented with the original bat-and-ball problem ("a bat and a ball cost $1.10 together …"). Participants were asked to indicate if they were familiar with the problem and to enter the correct response. Although it has been shown that prior exposure to the standard bat-and-ball has little impact on people's performance with the type of content modified versions we adopted (e.g., Chandler, Mueller, & Paolacci, 2014; but see also Meyer, Zhou, & Frederick, 2018), we wanted to eliminate the possibility that prior knowledge about the correct solution biased our results. Therefore, we decided to discard all data from participants who indicated they had seen the original bat-and-ball problem before and knew the correct response. This was the case for 44 participants (19% of total sample). Data of the 187 remaining participants (146 female, Mean age = 22.9 years, SD = 3.6 years) was entered into the analyses.

In addition, for all our latency based analyses in the present studies we a priori decided to discard trials with response latencies more than three standard deviations above the mean. In Study 1 this was the case for 4 trials. Due to a software problem latency and second guess data for one participant was missing and could not be analyzed. Second guess data on one additional trial was missing and could not be analyzed.

Results and discussion

Accuracy. Table 1 gives an overview of the findings. In line with previous studies, the vast majority of participants failed to solve the conflict version of the bat-and-ball problem correctly. Accuracy on the conflict problems only reached 27%. As expected, on the no-conflict control problems accuracy was almost at ceiling with 98% correct responses. Throughout the study we used mixed-effect models approach to analyze our data (accuracy and other), where we always entered the random intercept of subjects in the models. Mixed-effect logistic regression models showed that the accuracy difference on the conflict and no-conflict problems was significant, $\chi^2(1) = 236.77$, $p < 0.0001$, $b = 4.8$.

Conflict detection findings. In addition to participants' response to the bat-and-ball problems we also recorded their response latencies and response confidence. As in previous studies, these measures allow us to measure biased reasoners' error or conflict detection sensitivity. We therefore contrasted the response latencies (i.e., time elapsed between presentation of the problem and response submission) and confidence ratings for incorrectly solved conflict problems and correctly solved no-conflict problems (e.g. Johnson et al., 2016; Pennycook et al., 2015). Latencies were log transformed prior to analysis. Multilevel mixed-effect regression models showed that biased incorrect responders were less confident (i.e., a decrease of 14 percentage points) that their response was correct, $\chi^2(1) = 39.1$, $p < 0.0001$, $b = 13.35$, and needed more time (9.01 s increase) to make a decision, $\chi^2(1) = 45.5$, $p < 0.0001$, $b = -0.2$, when answering conflict vs no-conflict problems. This replicates previous findings that already indicated that biased reasoners on the bat-and-ball problem show conflict sensitivity (e.g., De Neys et al., 2013; Gangemi et al., 2014; Johnson et al., 2016)

Establishing the presence of a conflict signal is critical because it allows us to interpret the second guess findings unequivocally. As we noted, if biased reasoners have a non-specific conflict signal we expect them to select a second guess response randomly. However, such random selection might also occur if reasoners simply fail to detect conflict. Our conflict detection findings eliminate this latter possibility. Given that there is independent latency and

confidence evidence for the presence of a conflict signal, a random distribution of second guess responses would point to the non-specific nature of this signal.

Finally, note that for the small group of reasoners who solved the conflict problems correctly, the latency and confidence data do not present a pure measure of conflict detection (i.e., they both detect *and* resolve the conflict, Johnson et al., 2016; Pennycook et al., 2015). Nevertheless, the interested reader can find an overview of the correct latency and confidence data in Table 1.

Second guess results. Table 2 gives an overview of the second guess response distributions. As the top row of the table suggests, biased reasoners' second guesses were not random, $\chi^2(3) = 64.9$, $p < 0.0001$. Reasoners who failed to solve the critical conflict problem correctly nevertheless predominantly selected the correct response as their second guess. The correct response option was the most frequently selected second guess (46.3%), followed by the high foil option (37.5%). The low foil (10.3%) and extreme high foil (5.9%) were rarely selected. In and by itself, this might indicate that the modal biased reasoner has a highly specific conflict signal. Although they opt for the intuitive "10 cents" response, they know there is an alternative "5 cents" response and preferably select this when second guessing. However, this conclusion only follows in so far as our second guess is a pure measure of the specificity of the alternative response that reasoners considered (see further).

Table 2 includes the second guess distributions for correct conflict and no-conflict responses. These distributions were also not-uniform (correct conflict, $\chi^2(3) = 63.8$, $p < 0.0001$; correct no-conflict, $\chi^2(3) = 86.6$, $p < 0.0001$). On the no-conflict problems where the initially selected intuitive "10 cent" answer is also correct, reasoners favor the second guess that is smaller and closest to their initial correct response (i.e., "5 cents" the same response that is correct on the conflict version). Correct responders on the conflict problems have a preference for the heuristically cued "10 cents" response as their second guess.

The selection of our second guess response options was based on a set of a priori rules aimed to minimize possible second guess response biases or confounds. However, given the results one can envisage further alternative confounds that might explain the observed

findings. For example, one may note that among the two second guess options that are smaller than the heuristic response (i.e., lower foil "1 cent" and correct response "5 cents"), the correct second guess option is always closest to the heuristic response. Now, imagine that reasoners have a medium specific conflict signal and simply detect that the correct response is smaller than "10 cents" without knowing that it is specifically "5 cents". A mere anchoring effect might imply that they will prefer the response closest to the cued "10 cents" answer that they initially generated. This could lead them to favor "5 cents" over the lower foil "1 cent". Hence, the preference for the correct second guess in Study 1 does not necessarily point to the specific nature of the conflict signal. Likewise, the two most frequently selected second guess options (i.e., correct and high foil) are closest to the numerical "mean" of the four options. Although our options were not presented on a visual scale and appeared in random order, a general tendency to pick a response near the "middle" or numerical mean would also favor a non-random selection of correct and high foil responses. Indeed, although the correct response was selected more frequently than the high foil among incorrect responders, the difference did not reach significance, $\chi^2(1) = 1.3$, $p = 0.26$. Hence, the dominance of the correct second guess might also result from a mere preference for "average" values. In sum, while the findings in Study 1 are suggestive, they are not conclusive. Study 2 was designed to draw clearer conclusions by increasing the number of second guess options and inclusion of additional validation problems.

STUDY 2

Method

Participants

In Study 2, we recruited 143 participants (80 female, Mean age = 33.9 years, SD = 12.1 years) on the online crowdsource platform Prolific Academic. They received £0.70 for their participation. Only native English speakers from the USA, Canada, UK, Australia, or New-Zealand were allowed to take part. A total of 38.5% of the participants reported high school as

highest completed educational level, while 58.7% reported having a post-secondary education degree (2.8% reported less than high school).

Material

The experiment started with the presentation of the same two reasoning problems as in Study 1. We will refer to these as the "core" problems. The procedure was similar to Study 1 except for the number of presented second guess options. In Study 2, two additional response options were added. An extra "close" low foil (i.e., heuristic response – 5 or – 10) and extra "close" high foil (heuristic response + 5 or + 10) that numerically fell in between the heuristic and correct second guess (or high foil). For the problem version that totaled to $4.40 (heuristic response = 40 cents) this resulted in the following second guess options:

- o 10 cents *(lower foil)*
- o 20 cents *(correct response)*
- o 30 cents *(close low foil)*
- o 50 cents (*close high foil)*
- o 60 cents *(high foil)*
- o 70 cents *(extreme foil)*

For the problem version that totaled to $3.30 (heuristic response = 30 cents) this resulted in the following second guess options:

- o 5 cents *(lower foil)*
- o 15 cents *(correct response)*
- o 25 cents *(close low foil)*
- o 35 cents (*close high foil)*
- o 45 cents *(high foil)*
- o 55 cents *(extreme foil)*

Hence, the inclusion of the "close" foil guaranteed that any possible tendency of reasoners with a medium specific conflict signal to select the response below but closest to the heuristic response should no longer favor selection of the correct response. In addition, more response options should make it easier to identify a "mean" or "middle" response bias to randomly select one of the two options closest to the numerical mean.

As in Study 1, the second guess options appeared in a random order. The same second guess options were presented for the no-conflict problems. For reasoners who solved the conflict problem correctly, the correct second guess option was again replaced with the heuristic response. Both decision latencies and response confidence were recorded to measure reasoners' conflict detection sensitivity.

Additional validation problems. After participants had completed the first two "core" problems they were presented with a block of three additional bat-and-ball like problems. We will refer to these as the "*validation*" problems. Key feature is that for these problems we used a different second guess elicitation to help us validate the findings. As before, participants were first asked to solve the problem and enter their answer. However, next we simply asked them to indicate whether their second guess was smaller or larger than the heuristically cued response. For example, if the heuristic response was "60", the second guess question would read:

> Imagine that your answer to the problem you just solved doesn't turn out to be right.
> Do you think that the correct response is smaller or larger than 60?
> o smaller than 60
> o larger than 60

After participants had made a selection, they were also asked to enter the exact numerical second guess they had in mind:

> Can you also type down the precise answer you would give as your second guess? Even if you're not sure, just give your best guess.

Our rationale was that this structured second guess elicitation would give us a more fine-grained indication of the specific alternative response that reasoners were considering while minimizing the tendency to simply repeat the initial answer. Note further that on the three validation problems participants were not asked for a confidence rating to minimize the

possibility that the entered confidence rating would prime or bias the numerical second guess estimation.

The three validation problems had the same basic structure as the two bat-and-ball "core" problems but used further modified content adopted from Trouche (2016; see also Mata et al., 2017). Instead of listing the price of two goods, they referred to a different unit (e.g., number). Here are the three content materials we used:

> 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? *(correct response = 25; heuristic response = 50)*
>
> An apple and an orange weigh 160 grams altogether. The apple weighs 100 grams more than the orange. How much does the orange weigh? *(correct response = 30; heuristic response = 60)*
>
> Altogether, a book and a magazine have 270 pages. The book has 200 pages more than the magazine. How many pages does the magazine have? *(correct response = 35; heuristic response = 70)*

Among the three validation problems that participants solved were two conflict versions and one no-conflict version. Problems were presented in random order and for each participant it was randomly determined which content material was used for the conflict and no-conflict problems.

Exclusion criterion. At the end of the experiment participants answered standard demographic questions and were presented with the original bat-and-ball problem ("a bat and a ball cost $1.10 together …"). As in Study 1, we discarded all data from participants who indicated they had seen the original bat-and-ball problem before and knew the correct response. This was the case for 42 participants (29.4% of total sample). Data of the 101 remaining participants (61 female, Mean age = 34.1 years, SD = 12.3 years) was entered into the analyses.

As in Study 1, for all our latency based analyses trials with response latencies more than three standard deviations above the mean were a priori discarded. This was the case for 11 trials (2.2% of total) in Study 2. Due to technical problems, one second guess trial was not correctly recorded and excluded from the analysis.

Results and discussion

Participants first solved the two "core" bat-and-ball problems from Study 1 with multiple choice second guess format and afterwards they solved three additional validation problems. We start by presenting the full results for the core problems and afterwards move on to the validation problems.

Accuracy and conflict detection core problems. Table 1 (middle panel) shows the results. In line with previous findings, the vast majority of Study 2 participants failed to solve the conflict version of the bat-and-ball problem (17.8% mean accuracy) whereas accuracy on the no-conflict control version was almost at ceiling (98% mean accuracy). Mixed-effect logistic regression models showed that this difference was significant, $\chi^2(1) = 160.6$, $p < 0.0001$, $b = 5.4$. With respect to the conflict detection analysis we again contrasted the response latencies (i.e., time elapsed between presentation of the problem and response submission) and confidence ratings for incorrectly solved conflict problems and correctly solved no-conflict problems. Latencies were log transformed prior to analysis. Multilevel mixed-effect regression models showed that biased incorrect responders were less confident (i.e., a confidence decrease of 4.9 percentage points) that their response was correct, $\chi^2(1) = 5.7$, $p = 0.017$, $b = 4.7$, and needed more time (i.e., 5.1 s increase) to make a decision, $\chi^2(1) = 31.02$, $p < 0.0001$, $b = -0.4$, when answering conflict vs no-conflict problems. Although the effects are somewhat less pronounced, the pattern is fully consistent with the Study 1 results and previous findings. This confirms that biased reasoners on the bat-and-ball problem show conflict sensitivity.

Second guess results core problems. Table 3 gives an overview of the distribution of the second guess choices in Study 2. The top row shows the second guesses of the incorrect responders on the conflict problems. The pattern differed from what would be expected by chance alone, $\chi^2(5) = 58.3$, $p < 0.0001$. The dominant category was the lower close foil (e.g., "8 cents") with a selection frequency of 43.4%. This was followed by the higher close foil (e.g., "12 cents", 25.3%). The selection rate of these two second guesses significantly differed from each

other, $\chi^2(1) = 3.95$, $p = 0.047$, suggesting that it does not result from an uninformed general tendency to randomly pick either one of the "middle" options. There is a clear preference for the lower second guess that is closest to the heuristic response. The correct second guess response was selected by 14.5% of incorrect responders. Hence, this suggests that the dominance of the correct response in Study 1 simply resulted from the fact that it was closest to the heuristic response. Taken together, these findings indicate that the average biased reasoner has a medium specific conflict signal. People detect that the heuristic response is too high but do typically not realize that the correct response is precisely "5 cents". This claim will be further supported by the second guess findings on the validation problems.

For completeness, note that the second guess distributions for correct no-conflict and conflict responses were also not-uniform (correct conflict, $\chi^2(5) = 24.3$, $p < 0.001$; correct no-conflict, $\chi^2(5) = 94.03$, $p < 0.0001$). On the no-conflict problems reasoners favor the second guess that was smaller and closest to their initial correct response (e.g., "8 cents"). Correct responders on the conflict problems have a preference for the heuristically cued "10 cents" response as their dominant second guess.

Validation problems. Average accuracy on the conflict versions of the validation problems was 29.2% (SD = 45.6%) and 97% (SD = 17.1%) for the no-conflict problems, mixed-effect logistic regression, χ^2 (1) = 238.2, $p < 0.0001$, $b = 17.1$. But the critical question concerns the second guess results. Table 4 presents a full overview. Given that there were three different item contents (with different numerical heuristic and correct answer values) we present both the results for each item separately (top half) and results averaged over items (bottom half). Key observation is that the vast majority of incorrect responders on the conflict problems indicate that they prefer a second guess below the heuristic response: On average the frequency of this second guess choice reached 78.3%, which significantly differs from chance, $\chi^2(1) = 45.9$, $p < 0.0001$. This trend is observed on each individual item. However, when people who opted for the "below heuristic" second guess are subsequently asked to estimate the correct response, the exact correct response ("5 cents") is virtually never generated (5.3% of

cases). As the columns with deviations[3] show, biased reasoners on the conflict problems in the "below the heuristic" group generate a second guess estimate that is indeed smaller but close to the heuristic response (average deviation from heuristic response = -6.5, average deviation from correct response = 23.5). Correct conflict responders also predominantly opt for the "below the heuristic" second guess option (98% of cases) but subsequently give estimates that are very close to the correct response (average deviation from heuristic = -28.7, average deviation from correct response = 1.5). The box plots in Figure 1 show the actual distribution of second guess estimates for each of our three item contents to illustrate the findings.

Taken together, these results indicate that biased reasoners have a medium specific conflict signal. If reasoners would have a non-specific conflict signal (or simply fail to detect conflict), they should have displayed an equal preference[4] for a second guess below or above the heuristic response. If reasoners had a specific conflict signal and knew that their "10 cents" response conflicted with an alternative "5 cents" response, we would expect them to either generate this alternative response as their second guess or at least give an estimate that was closer to this response than to the incorrect heuristic response.

Individual differences. The overall pattern of second guess preferences give us an indication of the type of second guess that most biased reasoners prefer. This can inform us about the specificity of the conflict signal of the modal or typical biased reasoner which is our main interest in the present study. But obviously, there are individual differences here. Not all biased reasoners show the same type of conflict specificity. In this section we explore whether such individual differences in the specificity of the conflict signal might be linked to a different level or likelihood of conflict detection (Pennycook et al., 2015). Hence, we look at the size of the measured conflict detection effects for biased reasoners with different second guess estimation preferences.

[3] We excluded a total of 3 outlying responses (1% of total) from the calculations of mean estimates and deviations because they stated an extremely high number that was higher than the total of the units in a given item, namely "500", "569" and "275". These estimates lay respectively more than 15 times ("569" and "275") and 1 time ("500") above their respective interquartile range.

[4] Given that the number of possible natural numbers above the heuristic is infinite, a non-specific responder might even be expected to maximize her changes by opting for the "above the heuristic" answer.

We focus first on the "below heuristic" or "above heuristic" second guess choice on the validation problems as a grouping criterion. Reasoners with a medium specific or specific conflict signal should always opt for the "below heuristic" option. In theory, only participants with a non-specific conflict signal or participants who do not detect conflict can end up in the "above the heuristic" group. Contrasting the conflict detection sensitivity for these two groups allows us to test whether biased reasoners with a "less" specific conflict signal (i.e., second guess above the heuristic response) show less conflict sensitivity than those with a "more" specific signal (i.e., second guess below the heuristic response).

Note that we can use our below/above subgroup classification to look at conflict detection on the validation problems themselves (i.e., response time contrast for incorrectly solved conflict vs correctly solved no-conflict validation problems) and we can use the classification as a predictor to look at conflict detection effects on the initial core problems (i.e., both response time and confidence contrast on the core problems[5]). Next, we can look at conflict detection effects at a continuous (are individuals in group X more likely to show a *larger* detection effect?) and categorical (are individuals in group X more likely to show a conflict detection effect – i.e., show lower confidence on conflict vs no-conflict problems irrespective of the effect size?, e.g., Frey, Johnson, & De Neys, 2018) level. To test these associations we ran polychoric (when correlating two categorical variables) and polyserial (when correlating a categorical and a continuous variable) correlation analyses. We dummy coded whether an individual belonged to the "below" (0) or "above" (1) heuristic group and entered their corresponding conflict detection effect measure in the analysis. Table 5 shows the results. As the table indicates, although there was a slight trend towards a more pronounced detection effect in the "below heuristic" group (e.g., stronger confidence decrease and latency increase), the correlations were typically small and non-significant.

Next, we also looked at the correlation between the various conflict detection indexes and the precise second guess estimation value that participants gave on our final free-response validation question. More specifically, our idea was to test whether the extent of the deviation of one's second guess to the correct answer is related to the size of the conflict detection

[5] Each individual can contribute up to two observations in these analyses.

effect. Hence, instead of a categorical classification (i.e., second guess below/above heuristic) we test here whether biased reasoners who are "further off" are less/more likely to detect conflict. To test these associations we ran polyserial (when correlating a categorical and a continuous variable) and Pearson (when correlating two continuous variables) correlation analyses. As Table 5 shows, the general pattern again pointed to weak and non-significant associations.

Taken together, these exploratory results indicate that individual differences in the specificity of the error signal have little impact on the extent or likelihood of conflict detection per se. Individuals with a more and less specific conflict signal seem equally likely to show conflict detection effects.

STUDY 3

In Study 1 and 2 we opted for a multiple choice and structured second guess elicitation method. In Study 3 we experiment with an unstructured, free-response format. Clearly, any method has potential advantages and disadvantages. As we noted in Study 1 and 2, with a free-response format we risk that some participants will repeat their initial response which renders their second guess choice uninterpretable for our current purposes. On the other hand, an unstructured free-response format avoids any potential cueing from the response options that are provided with a multiple choice or semi-structured format. Hence, Study 3 allows us to test the robustness of our results. If we were to establish that most biased reasoners spontaneously generate a second guess below the heuristic response, this would provide additional validation for the Study 1 and 2 findings. In addition, the study allowed us to optimize our design to look more closely at the individual difference question and test our exploratory Study 2 findings.

Method

Participants

In Study 3, we recruited 140 participants (95 female, Mean age = 35.4 years, SD = 10.9 years) on the online crowdsource platform Prolific Academic. They received £0.50 for their

participation. Only native English speakers from the USA, Canada, UK, Australia, or New-Zealand were allowed to take part. A total of 56.4% of the participants reported high school as highest completed educational level, while 42.2% reported having a post-secondary education degree (1.4% reported less than high school).

Material

In Study 3 we used an unstructured free-response second guess elicitation. After participants had solved a problem they were presented with the following second guess question which appeared on a new page:

> Second guess:
> Imagine that your answer to the problem you just solved doesn't turn out to be right. You're allowed to make a second guess. Even if you're not sure, just give your best guess. Simply make sure to pick an answer that is different from your first answer.
>
> Please type your exact second guess bellow: "

Participants then entered their numerical estimate in a response box bellow the question.

To optimize the measurement of individual differences in conflict detection sensitivity each participant solved four problems (two conflict and two no-conflict). On all four problems response latencies and confidence were recorded. The item format was based on the modified bat-and-ball problems we used for the validation problems in Study 2. Instead of listing the price of two goods, the items referred to the number of different goods. Here are the specific contents we used for the conflict and no-conflict versions:

> Conflict items:
>
> Altogether, a book and a magazine have 260 pages
> The book has 200 pages more than the magazine
> How many pages does the magazine have?
>
> In a school there are 160 boys and girls in total
> There are 100 more boys than girls
> How many girls are there?

No-conflict items:

On a shelf there are 190 Pepsi and Coke bottles in total
There are 100 Pepsi bottles on the shelf
How many Coke bottles are there?

In a shop there are 290 PCs and Macs altogether
There are 200 PCs in the shop
How many Macs are there?

To optimize the individual differences analysis we presented the exact same versions to all participants (i.e., content and conflict status were not counterbalanced). So all participants solved the same two conflict and no-conflict items. This guarantees that any possible inter-item variability cannot bias the individual differences analyses. Moreover, to facilitate analysis of the second guess estimation deviation, the two conflict and two no-conflict problems had the same numerical value for the correct and heuristic response. Hence, for the conflict items values for the correct and heuristic response were always 30 units and 60 units, respectively. For the no-conflict items the correct and heuristic value was 90 units.

We reasoned that the use of an unstructured, free-response second guess format in Study 3 - in combination with the more structured formats we opted for in Study 1 and 2 - will give us the most general test of our hypothesis. However, as we noted in Study 1 and 2, the open-ended nature of the response format might imply that some participants will be tempted to repeat their initial response. In these cases, the second guess is not informative and will need to be discarded from the analyses. Furthermore, in theory, the fact that participants give a confidence rating before the second guess in Study 3 might prime and bias the second guess estimation. However, note that Study 1 and 2 established that even on the conflict problems average confidence ratings were fairly high (+80%). By selecting a heuristic value for our conflict problems (i.e., 60) that was considerably smaller, we minimized the possibility that the confidence rating would prime selection of a second guess below the heuristic. Hence, if anything, the confidence rating in Study 3 will work against our hypothesis that most biased reasoners give a second guess that is smaller than the heuristic response. Finally, one might note a possible problem with the latency based conflict detection measure in Study 1 and 2. Because the no-conflict items omitted the critical “more than” statement, they were also

slightly shorter. Hence, possible longer reading times might drive the decision latencies upwards. In Study 3 we added non-critical context words to the no-conflict versions so as to make sure that the average number of words (i.e., 24) in the conflict and no-conflict problems was similar.

Exclusion criterion. At the end of the experiment participants answered standard demographic questions and were presented with the original bat-and-ball problem ("a bat and a ball cost $1.10 together …"). As in Study 1 and 2, we discarded all data from participants who indicated they had seen the original bat-and-ball problem before and knew the correct response. This was the case for 20 participants (14.3% of total sample). Data of the 120 remaining participants (85 female, Mean age = 35.8 years, SD = 10.8 years) was entered into the analyses.

As in Study 1 and 2, for all our latency based analyses, we a priori decided to discard trials with response latencies more than three standard deviations above the mean (7 trials were discarded).

RESULTS

Accuracy. As Table 1 (bottom rows) indicates, accuracy findings were consistent with the Study 1 and 2 results and previous studies. Participants typically failed to solve the conflict problems correctly (average accuracy = 23.8%) whereas their performance on the no-conflict problems was at ceiling with an average of 97.1% correct responses. Mixed-effect logistic regression models showed that this accuracy difference was significant, χ^2 (1) = 384.3, p < 0.0001, b = 14.04.

Conflict detection findings. As before, we again contrasted the response latencies and confidence ratings for incorrectly solved conflict problems and correctly solved no-conflict problems. Latencies were log transformed prior to analysis. Multilevel mixed-effect regression models showed that biased incorrect responders were less confident (i.e., a 4.4 percentage point decrease) that their no-conflict response was correct, χ^2(1) = 13, p = 0.0003, b = 4.7, and

needed more time (i.e., 5.3 s increase) to make a decision, $\chi^2(1) = 48.4$, $p < 0.0001$, $b = -0.16$, when answering conflict vs no-conflict problems. These results again confirm our previous findings and indicate that biased bat-and-ball reasoners demonstrate conflict sensitivity.

Second guess results. In 17.8% of the trials (85 cases out of 480; 41 out of the 240 conflict, and 44 out of the 240 no-conflict problems) participants repeated their first, initial response on the second guess question. As we clarified in the method section, these trials were excluded from the second guess analyses. Table 6 presents a full overview of the analyzable second guess data. Given that the two conflict items had the same numerical heuristic and correct answer, we simply collapsed results over both item contents (idem for the no-conflict items). For comparison with Study 2 we also report the data separately for trials on which the second guess response was below or above the heuristic response value (i.e., 60 units).

A key observation is that a majority of incorrect responders on the conflict problems (53.4%) generate a second guess estimate below the heuristic response value. This proportion is smaller than what we observed with the forced choice validation format in Study 2 (i.e., 78.3%) and did not differ significantly from the proportion of trials in which an estimate above the heuristic value was generated, $\chi^2(1) = .68$, $p = .41$. However, this contrast is presented for completeness. One might argue that an equal distribution of responses under and above the heuristic value is a questionable null-hypothesis benchmark here. Values below the heuristic are confined to numbers between 0 and 59. Obviously, people can generate an infinite number of estimates above the heuristic value. Hence, if people had no specific insight into the nature of their error and were responding randomly, they should be more likely to end up with a response above the heuristic value. To test this statistically we can use the total sum that is defined in each item (i.e., 160 units and 260 units) as a practical upper limit for the range of possible second guesses[6]. This allows us to calculate the probability that one would end up with a second guess below the heuristic value for each of the two items when responding randomly

[6] Although in theory the upper limit is infinite, we reasoned that in practice it is unlikely that participants who read the preambles would generate a second guess that is higher than the total sum of the two objects. This total is the highest number that is explicitly listed in the problem. Indeed, there was only 1 case in Study 3 in which a higher second guess was observed. Clearly, although we find this assumption reasonable it remains speculative.

(i.e., 260 units total, 60/260 = .23; 160 units total, 60/160 = .375). Results showed that for each of our two conflict item contents we find that the generation rate of second guesses below the heuristic value is significantly higher than what would be expected by chance alone, $\chi^2(1) = 33.98$, $p < 0.0001$ (260 units, book and magazine content) and, $\chi^2(1) = 13.3$, $p = 0.0003$ (160 units, boys and girls content).

The fact that biased reasoners seem to display a tendency to spontaneously generate a second guess below the heuristic value supports the conclusion that they must at least have a medium specific conflict intuition and grasp that the correct response needs to be smaller than the heuristic answer they gave. But as Table 6 shows, even among those biased reasoners who generate a second guess below the heuristic response, the actual correct response is rarely generated (10.3% of cases). Furthermore, as the rows with deviations in Table 6 and the box plots in Figure 2 show, biased reasoners on the conflict problems in the "below heuristic" group typically generate a second guess estimate that is close to the heuristic response (average deviation from heuristic response = -10.2, average deviation from correct response = 19.8). These results are consistent with the Study 2 findings and argue against a highly specific nature of the conflict signal. Although most biased reasoners might realize that the correct response is smaller than the heuristic one, they typically do not know what the correct answer value is.

Further in line with the Study 2 findings, among correct conflict responders who give a second guess below the heuristic response, the generated estimates are close to the correct response (average deviation from heuristic = -36.1, average deviation from correct response = -6.1). However, in contrast with Study 2, opting for a second guess below the heuristic among correct responders was much rarer with the unstructured free response format in Study 3 (i.e., 37.7% of trials vs 98% in Study 2). One clear reason for this was that correct responders in Study 3 often generated the exact heuristic response as second guess (n = 22 or 41.5% of correct conflict trials). In Study 2 generation of the exact heuristic response was presumably discouraged because of the initial forced choice decision in the semi-structured elicitation (i.e., if I first indicate that the second guess is *smaller* than x, subsequently giving estimate x – instead of a value smaller than x - renders me mathematically inconsistent). The free response design in Study 3 sidestepped this complication. However, even if we combine the "below

heuristic" and "exact heuristic" group in Study 3, there are still considerably more correct responders who opt for a second guess above the heuristic response in Study 3 (20.8%) compared to Study 2 (2%). As we noted in the method section, it is possible that some of our design options (e.g., value of the heuristic response on no-conflict problems, confidence rating prior to second guess etc.) primed and pushed second guess estimates upwards.

Individual differences. In Study 3 we wanted to explore further whether individual differences in the specificity of the conflict signal are linked to a differential conflict detection sensitivity. Hence, as in Study 2, we again look at the size of the measured conflict detection effects for biased reasoners with different second guess estimations. In Study 3, all reasoners solved two conflict and no-conflict problems and response latencies and confidence data were recorded on all problems. For each biased participant we calculated the average conflict detection effect on the latency and confidence index (i.e., average response time and confidence contrast for incorrectly solved conflict vs correctly solved no-conflict problems). Next, we looked at the same set of correlations as in Study 2. That is, we considered both a continuous (i.e., size of the conflict detection effect) and categorical (i.e., whether or not the individual shows a conflict detection effect) conflict detection index and paired these with both the second guess estimate deviation from the correct response as well as a dummy coded categorical split-up into a below/above heuristic second guess classification[7]. Results are shown in Table 5. Although we again looked at a wide range of possible associations, correlations were overall weak and non-significant. This confirms the findings of Study 2 and indicates that possible individual differences in the specificity of the error signal have little impact on the extent or likelihood of conflict detection.

GENERAL DISCUSSION

In the present studies we introduced a second guess paradigm to test the specificity of error detection in the bat-and-ball problem. The results of three studies in which we used a

[7] Rare trials (n = 5) in which incorrect responders who generated an initial response different from the heuristic response subsequently gave the heuristic response as second guess were included in the above heuristic group.

range of second guess elicitation methods suggest that the average biased reasoner has a medium-specific error signal. When asked to make a second guess, participants predominantly give a second guess that is smaller than the heuristic "10 cents" response. This argues against the non-specific nature of the error signal. However, at the same time biased participants rarely select or generate the actual correct response and their second guesses remain close to the heuristic response. This argues against a highly specific nature of the error or conflict signal. Hence, biased reasoners who fail to solve the bat-and-ball problem correctly do not know what the correct answer is, but they at least seem to grasp that it needs to be smaller than the salient "10 cent" answer they gave into.

The key interest in the present study concerned the dominant second guess pattern among biased reasoners. This allows us to draw conclusions about the modal biased reasoner. However, it is important to bear in mind that there were individual differences. Some biased responders showed evidence of a non-specific (i.e., second guesses above the heuristic value) and - in a very limited number of cases - even a highly specific conflict signal (i.e., correct response as second guess). Hence, it is important to not forget that there are individual exceptions to the dominant pattern. We also tested whether these individual differences in the specificity of the error signal were associated with a differential conflict sensitivity. Our results indicated that this was not the case. Biased reasoners with a more specific and less specific conflict signal showed similar latency and confidence-based conflict detection effects. At a first pass, it might seem surprising that a "better" (i.e., more precise) signal, does not result in stronger effects. However, here it needs to be considered that precisely the absence of insight might render the processing doubt more pronounced (e.g., see Szollosi et al., 2017). Put differently, detecting that there is a problem with your answer without knowing what this problem is might be more unsettling than having at least partial insight into why the answer is problematic. This might blur the contrast between the different groups in the present approach[8].

[8] As one reviewer noted, an additional problem is that there might be individual differences in the expression of one's subjective conflict experience. For example, the same amount of subjective conflict experience might result in a differential slowing down/confidence decrease in subject x and y and vice versa.

In our three studies we tried to avoid any general potential priming or response bias confounds. Nevertheless, a critic could point to further potential confounds. For example, one might argue that the second guess question gives rise to a general "backward reasoning" confound. That is, people might only start to question their heuristic response during the second guess stage when they are explicitly alerted to the fact that their answer might be wrong. It would only be because of additional deliberation at this point that people realize that the "10 cents" answer is incorrect and that the correct answer needs to be smaller. However, our data argue against this account. First, our conflict detection findings indicate that participants already question their heuristic response during the first response stage. Second, we also analyzed the second guess response latencies across our three studies. If biased reasoners who gave a second guess below the heuristic "10 cents" only did so because the second guess cue helped them to deliberate further, then they should take longer to select a second guess option in comparison to biased reasoners who selected a second guess above the heuristic and presumably failed to engage in this additional deliberation. However, our results across the pooled Study 1-3 data clearly indicate that "below the heuristic" responders do not take more time to make their second guess decision (mean below group latency = 11.01 s, SD = 2.5; mean above group latency = 11.72 s, SD = 2.32, $\chi^2(1) = .48$, $p = .49$).

Another possible concern is that in all our studies it was always the case that the correct response ("5 cents") was smaller than the heuristic response ("10 cents"). Hence, a straightforward alternative explanation for our second guess results might be that reasoners simply have a general tendency towards second guesses that are smaller than their first answer. We believe that such a general confound is unlikely precisely because the heuristic response in the bat-and-ball problem is already at the lower end of the range of possible values (e.g., 10 cents vs total of 110 cents). Although we would expect that in the absence of any insight about the correct response people who give an initial response near the high end of a scale will indeed be more likely to pick a lower second guess, people whose initial response is near the lower end should be more likely to give a higher second guess. To test this directly we ran a very simple additional control study. In the study we presented a convenience sample of Hungarian university students (n = 63) with a problem for which they could not know the

correct answer. We told them “A computer has generated a random number between 1 and 200. Try to guess what this number is: ”. After they entered their first response, they were asked to enter a (self-generated) second guess following the same procedure as in Study 3. Results showed that participants who gave a first guess in the higher range (equal to or above the median first guess of 100) were more likely to give a second guess below their first guess (76% of cases). However, participants who gave a first guess in the lower range (below the median of 100) typically gave a second guess *above* their first guess (72% of cases). This argues against a general “lower second guess tendency” confound. If anything, in the absence of any knowledge about the correct response, our “10 cents” responders should have been much more likely to give a higher second guess.

Finally, one might also note that in our analyses we did not distinguish between incorrect conflict responders who gave the heuristic “10 cents” response and those who gave an incorrect non-heuristic response (e.g., “15 cents”) as their initial answer. Overall, non-heuristic incorrect responses are rare (i.e., across our three studies this amounted to 5.7% of incorrect conflict responses). As in previous conflict detection studies with the bat-and-ball problem (e.g., Bago & De Neys, 2018; Frey et al., 2018; Johnson et al., 2016) we refrained from a systematic discarding of non-heuristic incorrect responses. Nevertheless, one might wonder whether the non-heuristic responders distort the second guess findings. To get a general indication of this issue we calculated—across our three main studies—the proportion of second guesses below the heuristic response for incorrect initial conflict answers overall (61.17%), heuristic incorrect initial conflict answers only (61.75%), and non-heuristic incorrect initial conflict answers only (51.72%). The virtually identical figure (61.17% and 61.75%) for overall incorrect trials and heuristic incorrect trials indicates that inclusion of the rare non-heuristic incorrect trials does not distort the second guess results. However, the lower proportion of “below heuristic” second guesses among non-heuristic incorrect responders does suggest that this small group of a-typical incorrect responders might have a less specific conflict signal. Hence, if anything one might argue that our conclusions with respect to biased responders’ partial error insight would have been even stronger when restricted to the dominant heuristic “10 cents” incorrect responders.

Our results have some important theoretical implications. The bat-and-ball problem has been taken as the prototypical example of a "corrective" dual process view (Kahneman, 2011; Evans & Stanovich, 2013). This view entails that people are biased precisely because they fail to detect that their intuitively cued initial response is incorrect. This can lead to the characterization of biased reasoners as blind heuristic thinkers who fail to consider the most elementary logico-mathematical considerations (Marewski & Hoffrage, 2015). The findings from the conflict detection literature that pointed to people's error sensitivity started to question this view (e.g., Ball, Thompson, & Stupple, 2017; De Neys et al., 2013; Johnson et al., 2017; Pennycook et al., 2015). The present findings strengthen this questioning: Not only do people detect that their heuristic answer is problematic, their second guesses indicate that they are sufficiently mathematically savvy to at least partially figure out why. That is, people do grasp that if two items cost $1.10 in total and one cost a dollar more than the other, the other needs to cost less than 10 cents. In other words, although we might not manage to solve "X + Y = 1.10, Y = 1 + X, Solve for X" precisely, we do not fail to realize that "X < 10". Hence, people are more knowledgeable about elementary logico-mathematical principles than the classic dual process and heuristics and biases literatures suggest.

The fact that reasoners have some partial insight into the nature of their error raises an interesting question and possible counter-argument. A critic might argue that if - pace the classic dual process view - biased reasoners indeed know that the correct answer needs to be smaller than "10 cents", then why do they still give the "10 cents" response as their initial answer in the first place? Here it is important to keep in mind that that we should not simply assume that knowing that a response is problematic suffices to refrain from giving it (e.g., De Neys & Bonnefon, 2013). There are various instances in which humans display similar behavior. For example, many of us keep on smoking although we know perfectly well that it is lethal. Likewise, a blackjack player might know that the odds tell him to stand but nevertheless opt to hit another card (Walco & Risen, 2017). Risen (2016) has referred to this phenomena as "acquiescence", cases in which we behave against our better judgment. As Risen argued, traditional dual process theories as they have been put forward by Evans and Stanovich (2013) or Kahneman (2011) typically couple error detection and correction. As we noted above, the

core idea is that you fail to give the correct response because you did not notice it was incorrect - otherwise you would have corrected it. Risen suggested that the acquiescence phenomenon indicates that error detection and correction need to be decoupled. Error detection does not imply correction and a lack of correction does not imply a lack of detection. The present findings underscore this same point on the very task that has been considered the paradigmatic example of the standard view.

It has been argued that the intuitive heuristic response or behavior typically wins in case of conflict because it is stronger and more salient (i.e., has a higher activation strength, e.g., Bago & De Neys, 2017, or is generate more fluently, e.g., Pennycook, 2017; Thompson & Johnson, 2014) than the alternative correct insight. Interestingly, our second guess results might help to identify a critical mediating factor in this process. One of the reasons for why the heuristic response might be stronger is exactly that the alternative insight is not precise enough. If people realize that the correct answer needs to be smaller than the heuristic response without knowing what it is precisely, they are basically forced to pick a value randomly among the possible candidates (i.e., some number between 1 and 9) if they want to avoid the “10 cents” response. While this might be a rational strategy, it is presumably not compelling to give a response based on an (albeit educated) guess. Hence, rather than to pick a random response among a number of candidates we typically opt for the response that was generated most fluently. However, when this response is no longer an option, we readily opt for the smaller response. This tentative account illustrates how the lack of specificity of the error signal might help to explain why the heuristically cued “10 cents” still dominates as first option.

In general, one could note that our results—as the initial conflict sensitivity findings (e.g., De Neys et al., 2013; De Neys & Glumicic, 2008)—underline the role of metacognitive processes in reasoning (e.g., Ackerman & Thompson, 2017). Metacognition refers to the processes that monitor our ongoing thought processes. Although the metacognition literature has traditionally focused on memorization and knowledge retrieval it has been recently stressed that metacognition is equally critical for more complex processes such as reasoning and problem solving (Ackerman & Thompson 2015, 2017; Thompson et al., 2011). One

component of this emerging "meta-reasoning" framework is the confidence in one's intuitive responses (i.e., the "Feeling of Rightness" or FOR, Thompson et al., 2011). The present work indicates that the decreased response confidence that is observed when people err on the bat-and-ball problem is based on partial insight into the nature of one's error. This suggests that the monitoring process during reasoning seems to have some minimal accuracy.

A related general question concerns the ultimate nature of people's partial error insight. Where does biased reasoners' "it's less than 10 cents" knowledge come from? Our study was not designed to answer this question but we speculate it results from educated adults' extensive math practice through years of elementary and secondary education. We do not find it unreasonable that this should allow one to at least implicitly grasp that "$1.10 + 10 cents > $1.10". Note that this does not imply that participants' insight results from explicit calculation or deliberation. Previous conflict detection work already indicated that reasoners' error sensitivity is intuitive in nature (i.e., it is also observed when deliberation is minimized under time pressure or secondary task load, e.g., Bago & De Neys, 2018; De Neys, 2012; Johnson et al., 2016). As one reviewer noted, one could even conceive a link to the Approximate Number System (ANS, Dehaene, 1992; Libertus, Odic, & Halberda, 2012; Gilmore, McCarth, & Spelke, 2007). The ANS is part of our broader intuitive number sense which allows us to rapidly estimate the number of objects in real world settings, compare these numerical estimates, and perform basic arithmetic operations over these gut sense representations (Libertus et al., 2012) We would not object to the suggestion that biased reasoners' partial error insight is linked to an intuitive number sense. However, this is clearly speculative and will need further direct testing in future work.

That being said, the key contribution of our second guess approach is that it established that the dominant initial selection of the heuristic response does not imply a complete lack of insight into its erroneous nature. We believe that this illustrates the potential of the second guess paradigm for conflict detection and dual process studies. We hope that the present paper can sever as a proof-of-principle and that the method will be more widely adopted in future work. Second guessing provides us with a simple and powerful tool to gain deeper insight into the nature of heuristic bias and conflict detection during higher-order reasoning. What it

indicates so far is that we are more knowledgeable about our errors than many have hitherto been willing to believe.

ACKNOWLEDGEMENTS

Bence Bago was supported by a fellowship from the Ecole des Neurosciences de Paris Ile-de-France and the Scientific Research Fund Flanders (FWO-Vlaanderen). Support through the ANR Labex IAST is gratefully acknowledged. This research was also supported by a research grant (DIAGNOR, ANR-16-CE28-0010-01) from the Agence National de la Recherche. We would like to thank Darren Frey, Eoin Travers, Gaelle Vallee-Tourangeau, and the Psychonomics Amsterdam 2018 audience for valuable feedback and suggestions. Raw data can be retrieved from https://osf.io/ak346/.

REFERENCES

Ackerman, R., & Thompson, V. A. (2015). Meta-reasoning: what can we learn from meta-memory?. In Feeney, A., & Thompson, V. A., (Eds), *Reasoning as memory* (pp. 164-182). New York, NY: Psychology Press.

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 607-617.

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.

Bago, B., & De Neys, W. (in press). The smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking and Reasoning*.

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

Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory & Cognition*, *38*(2), 186–196.

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. Blaisdell, L. Huber, & A. Young, *Rational Animals, Irrational Humans* (pp. 232–252). Tokyo: Keio University Press.

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.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1-2), 1-42.

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

De Neys, W. (Ed.). (2017). Dual Process Theory 2.0. Oxon, UK: Routledge.

De Neys, W., & Bonnefon, J.-F. (2013). The 'whys' and 'whens' of individual differences in thinking biases. *Trends in Cognitive Sciences*, *17*(4), 172–178.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS One*, *6*(1), e15954.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299.

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., Vartanian, O., & Goel, V. (2008). Smarter Than We Think When Our Brains Detect That We Are Biased. *Psychological Science*, *19*(5), 483–489.

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.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science*, *17*(5), 383–386.

Evans, J. S. B. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321–339.

Ferreira, M. B., Mata, A., Donkin, C., Sherman, S. J., & Ihmels, M. (2016). Analytic and heuristic processes in the detection and resolution of conflict. *Memory & Cognition*, *44*(7), 1050–1063.

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

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *The Quarterly Journal of Experimental Psychology, 71,* 1188-1208.

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—in search of a phenomenon. *Thinking & Reasoning*, *21*(4), 383–396.

Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2007). Symbolic arithmetic knowledge without instruction. *Nature*, *447*(7144), 589-591.

Gilovich, T., Medvec, V. H., & Chen, S. (1995). Commission, omission, and dissonance reduction: Coping with regret in the" Monty Hall" problem. *Personality and Social Psychology Bulletin*, *21*(2), 182–190.

Handley, S. J., & Trippas, D. (2015). Chapter Two-Dual Processes and the Interplay between Knowledge and Structure: A New Parallel Processing Model. *Psychology of Learning and Motivation*, *62*, 33–58.

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.

Koriat, A. (2017). Can People Identify “Deceptive” or “Misleading” Items that Tend to Produce Mostly Wrong Answers? *Journal of Behavioral Decision Making*, *30*(5), 1066-1077.

Libertus, M. E., Odic, D., & Halberda, J. (2012). Intuitive sense of number correlates with math scores on college-entrance examination. *Acta psychologica*, *141*(3), 373-379.

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.

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: an attentional account of reasoning errors. *Psychonomic Bulletin & Review*, *24*(6), 1980-1986.

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision Making, 13(3)*, 246-259.

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.

Oster, N., & Koesterich, R. (2013). Breaking Bad Behaviors: Understanding Investing Biases and How to Overcome Them. *IShares Market Perspectives*, 1–9.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (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., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 544–554.

Risen, J. L. (2016). Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions. *Psychological Review*, *123*(2), 182-207.

Simon, G., Lubin, A., Houdé, O., & De Neys, W. (2015). Anterior cingulate cortex and intuitive bias detection during number conservation. *Cognitive Neuroscience*, *6*(4), 158–168.

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.

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.

Stupple, E. J., Ball, L. J., Evans, J. S. B., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychology*, *23*(8), 931–941.

Stupple, E. J., Thompson, V. A., & Ball, L. J. (2017). Conflict and dual process theory: the case of belief bias. In W. De Neys, *Dual Process Theory 2.0* (pp. 108-128). Oxon, UK: Routledge.

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.

Szollosi, A., Bago, B., Szaszi, B., & Aczel, B. (2017). Exploring the determinants of confidence in the bat-and-ball problem. *Acta Psychologica*, *180*, 1–7.

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

Thompson, V. A., 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*.

Vartanian, O., Beatty, E. L., Smith, I., Blackler, K., Lam, Q., Forbes, S., & De Neys, W. (2018). The reflective mind: Examining individual differences in susceptibility to base rate neglect with fmri. *Journal of Cognitive Neuroscience, 30*(7), 1011-1022.

Villejoubert, G. (2009). Are representativeness judgments automatic and rapid? The effect of time pressure on the conjunction fallacy. In *Proceedings of the Annual Meeting of the Cognitive Science society* (Vol. 30, pp. 2980–2985). Cognitive Science Society.

Walco, D., & Risen, J. L. (2017). The empirical case for acquiescing to intuition. *Psychological science*, *28*(12), 1807-1820.

Table 1. Overview of accuracy and conflict detection findings. Number of trials (italics) and standard deviation between brackets.

		Conflict		No conflict	
Study	Measure	Correct	Incorrect	Correct	Incorrect
Study 1	Accuracy	27.3% (*51*)	72.7% (*136*)	97.9% (*183*)	2.1% (*4*)
	Conflict detection				
	Response time	69.4 s (2.4)	25.5 s (2.2)	16.9 s (1.9)	12.8 s (3.9)
	Confidence rating	96.2% (7.5)	81.3% (27.9)	95.2% (14.1)	100% (0)
Study 2	Accuracy	17.8% (*18*)	82.2% (*83*)	98% (*99*)	2% (*2*)
	Conflict detection				
	Response time	35.4 s (1.9)	15.2 s (2)	10.1 s (1.5)	9.4 s (1.4)
	Confidence rating	98.5% (3.3)	92.1% (20.4)	97% (12)	100% (0)
Study 3	Accuracy	23.8% (*57*)	76.2% (*183*)	97.1% (*233*)	2.9% (*7*)
	Conflict detection				
	Response time	32.6 s (2.1)	18.9 s (2.2)	12.8 s (1.8)	24.7 s (3.1)
	Confidence rating	96.4% (8.1)	93.3% (17.5)	97.7% (10.9)	84.3% (3.1)

Table 2. Second guess response distribution across the four options in Study 1. Number of trials between brackets.

	Second guesses			
Response type	Low foil ("1 cent"[+])	Correct[*] ("5 cents")	High foil ("15 cents")	Extreme foil ("25 cents")
Conflict incorrect	10.3% (14)	46.3% (63)	37.5% (51)	5.9% (8)
Conflict correct	28% (14)	70% (35)	2% (1)	-
No-conflict incorrect	25% (1)	50% (2)	25% (1)	-
No-conflict correct	11% (20)	48.6% (88)	34.3% (62)	6.1% (11)

Notes. [+]Second guess options are illustrated with numerical values based on the original bat-and-ball problem. [*]For correct conflict responses the correct second guess alternative was replaced with the heuristic response (e.g., "10 cents").

Table 3. Second guess response distribution for the core items across the six options in Study 2. Number of trials between brackets.

	Second guesses					
Response type	Low foil[+] ("1 cent")	Correct* ("5 cents")	Low close ("8 cents")	High close ("12 cents")	High foil ("15 cents")	Extreme foil ("25 cents")
Conflict incorrect	8.4% (7)	14.5% (12)	43.4% (36)	25.3% (21)	4.8% (4)	3.6% (3)
Conflict correct	11.8% (2)	47.1% (8)	41.2% (7)	-	-	-
No-conflict incorrect	-	50% (1)	-	-	50% (1)	-
No-Conflict correct	7.1% (7)	13.1% (13)	46.5% (46)	29.3% (29)	3% (3)	1% (1)

Notes. [+]Second guess options are illustrated with numerical values based on the original bat-and-ball problem.*For correct conflict responses the correct second guess alternative was replaced with the heuristic response (e.g., "10 cents").

Table 4. Overview of second guess findings (below/above heuristic forced choice and free response estimate) on the validation problems in Study 2. Data are shown for each of three different item contents. Number of trials between brackets.

			Second guess below heuristic					Second guess above heuristic				
Item	Version	Accuracy	Frequency	Mean estimate	Frequency correct estimate	Deviation Correct	Deviation Heuristic	Frequency	Mean estimate	Frequency correct estimate	Deviation Correct	Deviation Heuristic
Item 1:	Conflict	0	80% (35)	51.74	5.7% (2)	21.74	-8.26	20% (9)	64.67	0%	34.67	4.66
"Total		1	100% (22)	30.27	-	0.27	-29.73	-	-	-	-	-
of 160	No-	0	100% (1)	60	100% (1)	0	0	-	-	-	-	-
units"	conflict	1	74%(25)	54.26	-	-5.74	-5.74	26%(9)	54.96	-	-5.04	-5.04
Item 2:	Conflict	0	75%(38)	43.84	7.9% (3)	18.84	-6.16	25%(13)	75.54	0%	50.54	25.54
"Total		1	94%(17)	28.18	-	3.18	-21.82	6%(1)	24	-	-1	-26
of 250	No-	0	-	-	-	-	-	-	-	-	-	-
units"	conflict	1	75%(24)	43.79	-	-6.21	-6.21	25%(8)*	111	-	61	61
Item 3:	Conflict	0	81%(39)**	64.76	2.6% (1)	29.8	-5.24	19%(9)	92.78	0%	57.78	22.78
"Total		1	100% (19)	36.47	-	1.47	-33.53	-	-	-	-	-
of 270	No-	0	100% (2)	40	0%	-30	-30	-	-	-	-	-
units"	conflict	1	81%(26)***	60.88	-	25.88	-9.12	19%(6)	73.83	-	3.83	3.83
Average	Conflict	0	78.3%(112)	-	5.4% (6)	23.5	-6.5	21.7%(31)	-	0%	48.03	18.7
		1	98.3% (58)	-	-	1.5	-28.7	1.7% (1)	-	-	-1	-26
	No-	0	100% (3)	-	33% (1)	-20	-20	-	-	-	-	-
	conflict	1	76.5% (75)	-	-	-7.03	-7.03	23.5%(23)	-	-	18.4	18.4

Note. We excluded a total of 3 responses from the calculations of mean estimates and deviations, because they stated an extremely high number that was higher than the total of the units in a given item, namely "500"*, "569"** and "275"***. These estimates (** and ***) lay more than 15 times and respectively 1 time (*) above their respective interquartile range.

Table 5. Overview individual difference findings. Conflict detection effect size among incorrect responders as a function of second guess performance on validation problems. Number of analyzed trials are in brackets.

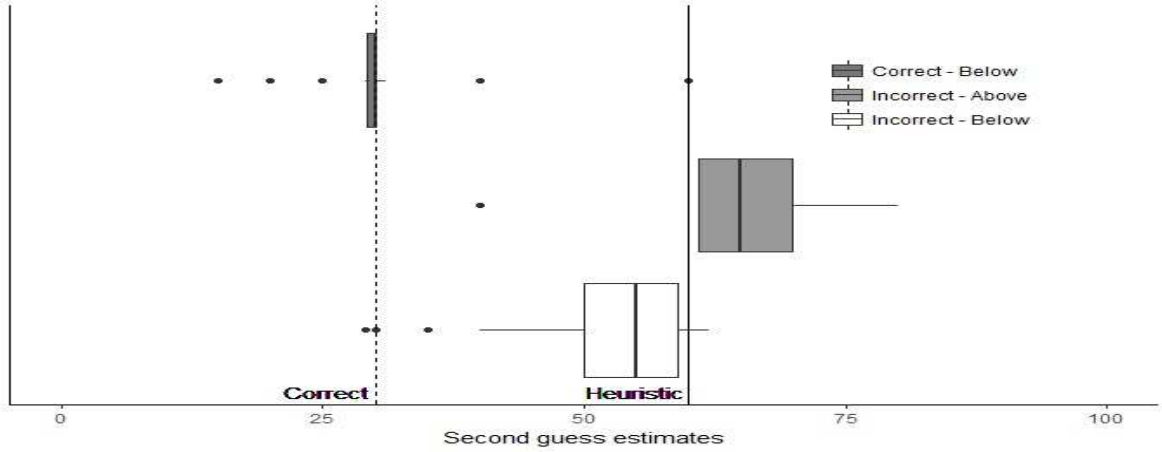
	Conflict detection measure[+]				
Second guess individual difference measure	Study 2 Core confidence	Study 2 Core latency	Study 2 Validation latency	Study 3 Confidence	Study 3 Latency
Average					
Below heuristic group	1.9 %	-4.5 s	-2.2 s	4.4%	-3.9 s
Above heuristic group	0.39 %	-3.8 s	-1.5 s	4.7%	-8.9 s
Correlations					
Below/above binary	0.15 (141)	0.03 (135)	-0.01 (136)	-0.05 (140)	-0.01 (138)
Below/above continuous	-0.1 (141)	0.09 (135)	0.17 (136)	0.02 (140)	-0.12(138)
Estimation binary	-0.11 (141)	-0.13 (135)	0.06 (136)	-0.03 (140)	0.09 (138)
Estimation continuous	0.02 (141)	0.09 (135)	0.01 (136)	0.02 (140)	-0.13 (140)

Notes. [+]Difference score: correct no-conflict problem minus incorrect conflict problem. Negative latency values and positive confidence values point to stronger detection effect.

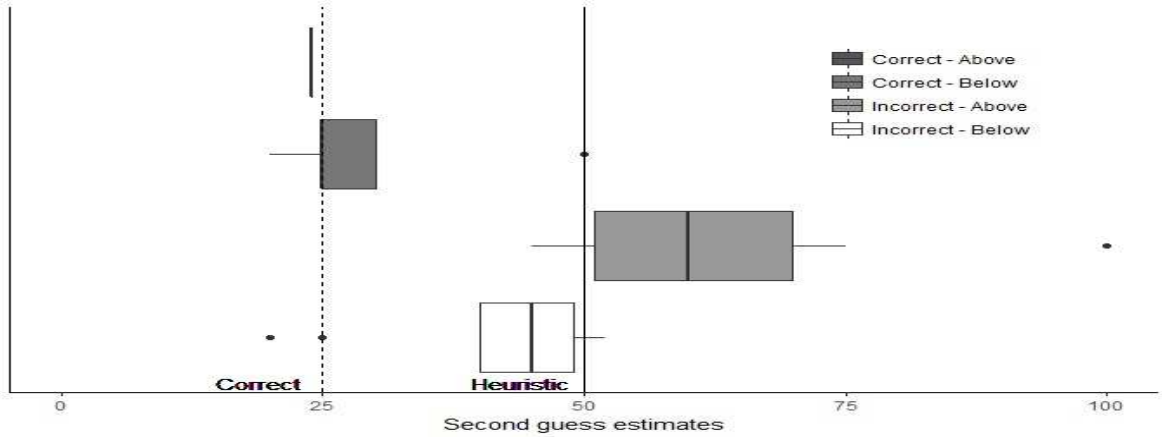
Table 6. Overview of average free response second guess estimation findings in Study 3. Number of trials (italics) and standard deviation between brackets.

	Conflict		No conflict	
Second guess estimate	Correct	Incorrect	Correct	Incorrect
Second guess below heuristic				
Frequency	37.7% (*20*)	53.4% (*78*)	48.4% (*93*)	25% (*1*)
Mean Estimate	23.9 (15.2)	49.8 (15.2)	61.4 (34.7)	0
Frequency correct estimate	-	10.3% (*8*)	-	-
Deviation correct	-6.1	19.8	-28.6	-90
Deviation heuristic	-36.1	-10.2	-28.6	-90
Second guess above heuristic				
Frequency	20.8% (*11*)	43.2% (*63*)	51.6% (*99*)	25% (*1*)
Mean Estimate	172.8 (53.6)	141.2 (61.7)	134.2 (60.5)	100
Frequency correct estimate	-	-	-	-
Deviation correct	142.8	111.2	44.2	10
Deviation heuristic	112.8	81.2	44.2	10
Second guess equals heuristic				
Frequency	41.5% (*22*)	3.4% (*5*)	(*0*)	50% (*2*)
Mean Estimate	60	60	-	90
Frequency correct estimate	0%	0%	-	100%
Deviation correct	-30	-30	-	0
Deviation heuristic	0	0	-	0

a. Item content 1 ("Total of 160 units")



b. Item content 2 ("Total of 250 units")



c. Item content 3 ("Total of 270 units")

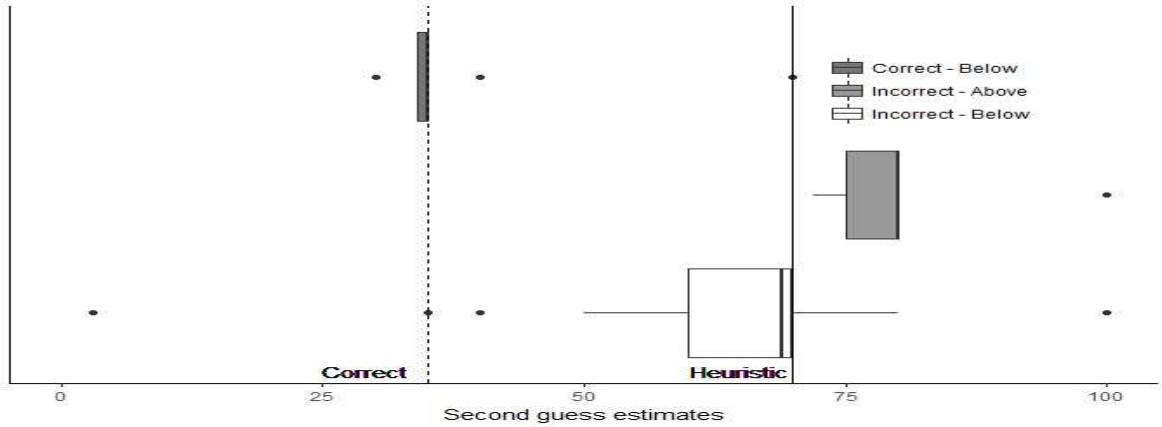
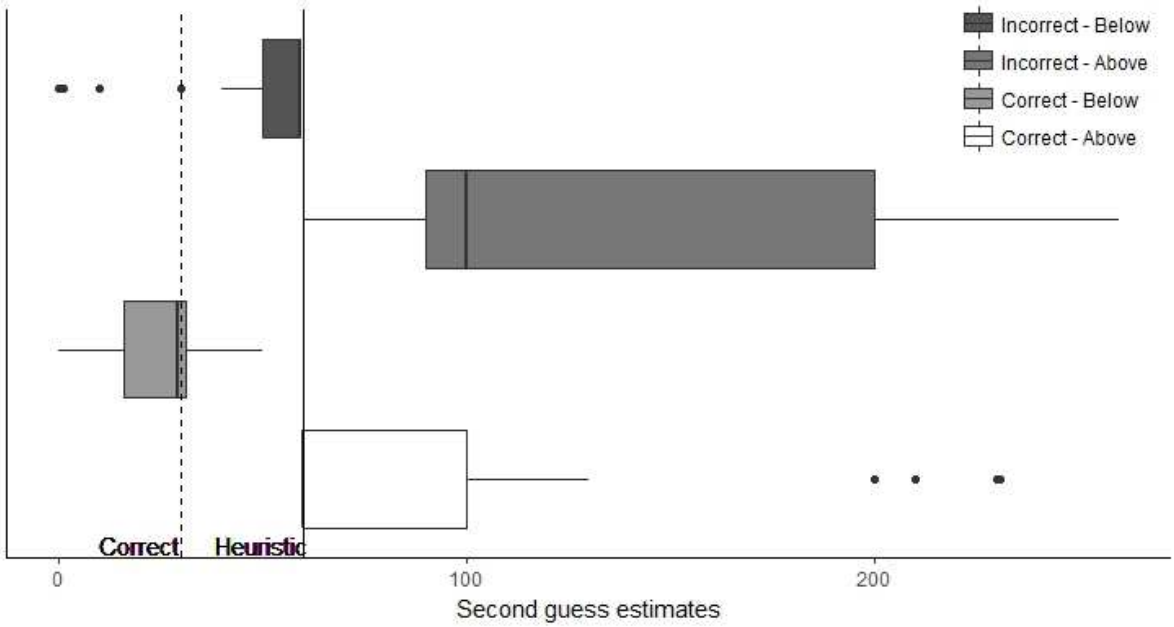


Figure 1. Box plots showing the distribution of second guess estimates on conflict validation trials for correct and incorrect responders who opted for a second guess below or above the heuristic response value in Study 2. Estimates are shown for each of the three different item contents we used (panel a., b., c.). Vertical dashed line shows value of correct response, solid line shows value of heuristic response. For the sake of presentation, values above 100 are not shown in these graphs, but they were taken into account for the calculation of the interquartile range.

a. Conflict items



b. No-conflict items

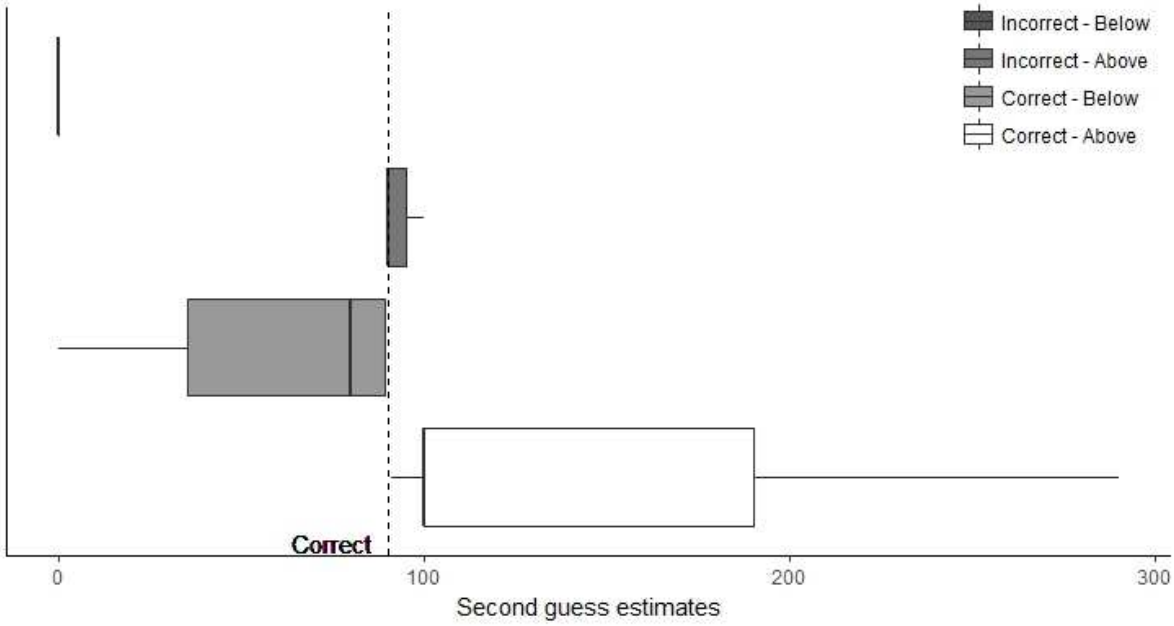


Figure 2. Box plots showing the distribution of second guess estimates on conflict (a) and no-conflict (b) trials for correct and incorrect responders in Study 3 who opted for a second guess below or above the heuristic response value. For ease of presentation trials in which the second guess equaled the heuristic response are included in the "above heuristic" group. Vertical dashed line shows value of correct response, solid line shows value of heuristic response.