Memory & Cognition, in press

THINK SLOW, THEN FAST: DOES REPEATED DELIBERATION BOOST CORRECT INTUITIVE RESPONDING?

Matthieu Raoelison[1], Marine Keime[2], Wim De Neys[1]

[1]Université de Paris, LaPsyDÉ, CNRS, F-75005 Paris, France
[2]University of Glasgow, Scotland

Abstract

Influential studies on human thinking with the popular two-response paradigm typically ask participants to continuously alternate between intuitive ("fast") and deliberate ("slow") responding. One concern is that repeated deliberation in these studies will artificially boost the intuitive, "fast" reasoning performance. A recent alternative two-block paradigm therefore advised to present all fast trials in one block before the slow trials were presented. Here we tested directly whether allowing people to repeatedly deliberate will boost their intuitive reasoning performance by manipulating the order of the fast and slow blocks. In each block participants solved variants of the bat-and-ball problem. Maximum response time in fast blocks was 4s and 25s in the slow blocks. One group solved the fast trials before the slow trials, a second group solved the slow trials first, and a third mixed group alternated between slow and fast trials. Results showed that the order factor did not affect accuracy on the fast trials. This indicates that repeated deliberation does not boost people's intuitive reasoning performance.

Keywords: Dual process; Reasoning; Decision-making; Two-response paradigm; Methodology

Introduction

Popular fast-and-slow dual-process theories have long conceived human reasoning as an interaction between a fast intuitive process and a slower deliberate reasoning process.[1] Although these theories have been very influential, their core assumptions are still debated (e.g., De Neys, 2017; De Neys & Pennycook, 2019; Evans, 2019; Evans & Stanovich, 2013). Recent discussions have been fueled by findings with the "two-response" paradigm (Thompson, Prowse Turner, & Pennycook, 2011). In this paradigm participants are asked to give two consecutive responses to a reasoning problem. First, they need to answer as fast as possible with the first response that comes to mind. Next, the same problem is presented again and participants can take all the time they want to deliberate before giving a final response.

Given that deliberation is believed to be more time-consuming than intuitive reasoning (Evans & Stanovich, 2013; Kahneman, 2011), the paradigm allows us to track how people's first intuitive hunch is modulated after subsequent deliberation. To make maximally sure that the first response is generated intuitively it typically needs to be generated under challenging time constraints (and/or instructions to give the first response that comes to mind, e.g., Newman, Gibb, & Thompson, 2017, and/or cognitive load, e.g., Bago & De Neys, 2017). Typical findings have been quite surprising. Traditional dual process models entail that sound reasoning in classic reasoning tasks requires deliberate correction of an incorrect intuition. Consider for example the infamous bat-and-ball problem:

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

For many people the first answer that intuitively pops up in mind is 10 cents. It is generally assumed that arriving at the correct response (5 cents) requires us to engage in slower and more demanding deliberate thinking and correct the intuitively generated 10 cents response (Frederick, 2005; Kahneman, 2011). Consequently, one would expect that in two-response studies, correct responding will typically be observed in the final response stage. However, results show that this is not the case. Two-response studies with the

[1]What might constitute an intuitive process is up to debate. In this article, we take a neutral stance and adopt an operational definition. Intuitive processes are faster and require fewer cognitive resources than deliberate processes (e.g., Kahneman, 2011).

bat-and-ball problem and other classic reasoning tasks show that correct responses are frequently generated in the initial response stage (Bago & De Neys, 2017, 2019b; Newman et al., 2017; Raoelison & De Neys, 2019; Thompson et al., 2011). Hence, good reasoners are not necessarily good at correcting erroneous intuitions, they simply seem to have more accurate intuitions (Bago & De Neys, 2017; Thompson, Pennycook, Trippas, & Evans, 2018). This forces us to revise our view on the nature of sound reasoning.

It is clear that the two-response paradigm findings can have far-stretching implications (De Neys & Pennycook, 2019). However, the conclusions only hold in so far as the paradigm itself is sound and does not introduce confounds. Although the paradigm has been validated in a number of studies (Bago & De Neys, 2017, 2019a, 2019b; Thompson et al., 2011), concerns remain (Markovits, de Chantal, Brisson, & Gagnon-St-Pierre, 2019; Rosas & Aguilar-Pardo, 2019). One key issue is that when multiple reasoning trials with the same or structurally similar problems are being presented (as is typically the case in reasoning studies), the repeated two-response deliberation stages might help people arrive at the correct solution. If one figures out the solution during the deliberation stage of a trial, then the next intuitive response on the subsequent trial, could benefit from the earlier deliberation and consequently would not reflect mere intuitive processing. Hence, intermittent deliberation, that is, constantly switching between intuitive and deliberate response stages, could artificially boost correct responding in the "intuitive" response stages. Consequently, the standard two-response paradigm would overestimate the prevalence of intuitive correct responding.

Recently, Markovits et al. (2019) presented an alternative paradigm in which participants were first presented with a full block of “fast” intuitive trials in which they always needed to answer with the first response that came to mind. Subsequently, participants were given a block of "slow" deliberate trials in which they were given all the time they wanted to solve the problem. This "two-block" design might give a purer measurement of people’s intuitive reasoning performance since it only allows participants to deliberate after all intuitive trials have been completed. However, Markovits et al. did not directly test for the potentially confounding impact of intermittent deliberation. Here we present a simple test by manipulating the order of the fast and slow blocks. This allows us to directly examine whether allowing people to deliberate will boost their intuitive reasoning performance. We had participants solve both a fast block and a slow one, each containing 50 bat-and-ball problem trials. Half of the participants started with a fast block, the other half with a slow block. If allowing people to repeatedly deliberate about a problem in the slow blocks enables them to learn the correct solution strategy, then we expect that we will observe more correct responses in the fast/intuitive block for those participants who first solved a block of slow/deliberative trials than for those who had to start with the fast/intuitive block.

For exploratory purposes we also included a condition in which participants continuously alternated between slow and fast blocks. In addition, we also introduced a number of methodological refinements to further optimize the two-block design. Our instructions were closely modeled after the original two-response bat-and-ball studies of Bago and De Neys (2019b). Furthermore, as in previous studies, we adopted a multiple-choice answer format but presented the problem question and response alternatives separately to minimize possible backward inferencing (i.e., the presented response options are used as a cue to guide/help the reasoning process, e.g., Bago, Raoelison, & De Neys, 2019, see methods).

1 Methods

1.1 Participants

We recruited 123 participants (79 female, Mean age = 34.9 years, SD = 12.9 years[2]) on Prolific Academic (www.prolific.ac). They were paid £5 per hour for their participation. Only native English speakers from Canada, Australia, New Zealand, the United States of America, or the United Kingdom were allowed to take part in the study. Among them, 48 reported high school as their highest level of education, while 73 had a higher education degree, and 2 reported less than high school as their highest educational level.

1.2 Material

The materials were taken from the study by Raoelison and De Neys (2019), who designed a total of 110 items. Of those, 50 items were variations of the bat-and-ball problem that had the same underlying structure as the original problem but different superficial

[2]Due to a technical error, we reported age as missing for 3 participants that were deemed to be above 100 years old.

item content (e.g., "In a company there are 150 men and women in total. There are 100 more men than women. How many women are there?"). Each problem specified two types of objects with different quantities instead of prices (e.g., see Bago & De Neys, 2019b; Mata, Ferreira, Voss, & Kollei, 2017). Each of the 50 problems featured unique content with a total amount that was a multiple of ten and ranged from 110 to 650. Each problem had four corresponding response options; the correct response ("5 cents" in the original bat-and-ball), the intuitively cued "heuristic" response ("10 cents" in the original bat-and-ball), and two foil options. Mathematically speaking, the correct equation to solve the original bat-and-ball problem is: \$1.00 + 2x = \$1.10, instead, people are thought to be intuitively using the "\$1.00 + x = \$1.10" equation to determine their response (Kahneman, 2011). We always used the latter equation to determine the "heuristic" answer option, and the former to determine the correct answer option for each problem. Following Bago and De Neys (2019b), the two foil options were always the sum of the correct and heuristic answer (e.g., "15 cents" in original bat-and-ball units) and their second greatest common divider (e.g., "1 cent" in original units). For each item, the four response options appeared in a randomly determined order. The following illustrates a full problem:

> *In a company there are 150 men and women in total.*
> *There are 100 more men than women. How many women are there?*

o 50

o 75

o 5

o 25

Note that the response options above are displayed for clarity. In the study, the response options were not displayed on the same screen but separately (see further). To avoid that the task would become too repetitive and to verify that participants stayed minimally engaged in the task there were also 50 control problems. In the standard bat-and-ball versions the intuitively cued "heuristic" response cues an answer that conflicts with the correct answer. In the "no-conflict" control problems, the heuristic intuition was made to cue the correct response option. This was achieved by deleting the critical relational "more than" statement (De Neys, Rossi, & Houdé, 2013; Travers, Rolison,

& Feeney, 2016). With the above example, a no-conflict problem version would look as follows:

In a company there are 150 men and women in total.
There are 100 men. How many women are there in this company?

o 5

o 50

o 25

o 75

In this case the intuitively cued "50" answer was also correct. We presented the same four answer options as for a corresponding standard conflict version. We added three words to the control problem question (e.g., "in this company") so that standard "conflict" and control "no-conflict" versions had roughly the same length. Given that the control items can be solved correctly on the basis of mere intuitive reasoning, we expected to see near-ceiling performance on the control items throughout, if participants were paying minimal attention to the task and refrain from mere random responding. Finally, in addition to the 50 conflict and the 50 no-conflict problems, there were also 20 filler problems in which participants simply had to add two quantities. For example:

In France, there are 50 million adults and 15 million children.
How many adults and children are there in total?

o 65 million

o 56 million

o 500 million

o 150 million

Note that in addition to the 10 original filler problems from Raoelison and De Neys (2019), we added 10 similar extra ones. The rationale behind the filler problems was that these would further help to render the task less repetitive and predictable. In total, participants had to solve 120 problems. The problems were grouped into ten blocks containing each 5 standard conflict problems, 5 control no-conflict problems, and 2 filler problems, presented in a randomized order. Participants could take a short break after completing each block. The conflict content was crossed across blocks, such that conflict problems in the first five blocks had their corresponding no-conflict versions in the last five blocks, and vice-versa. Hence, all items within a block had a different superficial item content.

The response options on each trial were always presented on a new screen after the problem premises and question (e.g., "In a company there are 150 men and women in total. There are 100 men. How many women are there in this company?") had been presented. Once the four response options appeared, the premises and question (which were presented line by line, see further) disappeared and participants had 2.5 s to move the mouse and click on their response. We reasoned that this procedure would force participants to calculate the answer themselves during the question presentation phase and minimize any potential backward inferencing (i.e., the answer options are used as possible solution cue, Bago et al., 2019), thus making it more similar to a free-response procedure.

Fast and slow trials. Half of the trials were "fast" trials in which participants were instructed to respond as fast as possible to each trial with the first intuitive response that came to mind. To ensure this, the problem was presented for maximally 4 s. The other half of the trials were slow trials in which participants were instructed they could take the time to deliberately reflect on the problem. The problem presentation duration was therefore extended to 25 seconds for slow trials. After the problem presentation, participants always had 2.5 seconds to enter their response. If participants were ready to enter a response before the maximum allotted presentation time had passed, they could advance to the response selection phase by clicking on a button labeled "Next". The fast and slow duration deadlines were inspired by the pilot work of Bago and De Neys (2019b) and Raoelison and De Neys (2019). Note that these studies used a traditional design in which question and response options were presented simultaneously. Nevertheless, they suggested that a 4 s deadline should put participants under considerable time-pressure whereas 25 s should give them ample time to reflect.

Each block of 12 trials always contained only fast or slow trials and participants were instructed at the beginning of each block which type of trials (fast or slow) they were going to get. Our critical manipulation concerned the order in which fast and slow blocks were presented. In the fast first condition, the first five blocks were fast blocks, followed by five slow blocks. This was reversed in the slow first condition. In addition, we created a "mixed" condition in which fast and slow blocks always alternated. Half of the participants in the mixed condition started with a fast block, the other half with a slow block. Participants were randomly allocated to the different conditions.

To minimize the possibility that the specific item content would confound the fast and slow first condition contrast, the content that was used for the fast trials in the fast first condition was used as content for the fast trials in the slow first condition for half of the participants and as material for the slow trials in the slow first condition for the other half (and vice-versa). The order of blocks in the mixed conditions was predetermined but the order of fast blocks and slow blocks in the fast first and the slow first conditions were randomized. For a graphical illustration of the full counterbalancing, see the Supplementary section (Figure S1).

1.3 Procedure

The experiment was run online on the Qualtrics platform. Participants were given the following general instructions:

> *Welcome to the experiment!*
>
> *Please read these instructions carefully!*
>
> *In this study, we are interested in how people solve problems when they're relying on their intuition (i.e., "fast" thinking) and how they solve them when they take the time to reflect on them. We will present you a set of 120 questions and a couple of practice questions. It will take about 30 minutes to complete and demands your full attention. You can only do this experiment once.*
>
> *We will always instruct you as to whether we want you to solve a problem "intuitively" or whether you can take your time to deliberate on your decision. For the intuitive trials, we want you to respond as fast as possible with the first answer that comes to mind. You don't need to think about it, just give the first response that comes to mind as quickly as possible. Be careful, you will only be given a few seconds to pick your answer! For the deliberative trials, you'll have the time to really reflect on the problem. Here we want you to think as deeply as possible before you give your answer.*

Please confirm below that you read these instructions carefully and then press the "Next" button.

Once they gave their consent, more specific instructions provided a more detailed explanation about the two types of trials:

We are going to start with a couple of practice problems.

For all problems, a fixation cross will appear first. Then, the first sentence of the problem is going to be presented for 2 seconds followed by the entire problem.

In the intuitive trials, we are interested in your initial, intuitive response. We want you to respond with the very first answer that comes to mind. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible. To assure this, a very short time limit was set when presenting the "intuitive" problem, which is going to be 4 seconds. If possible, we want you to answer even faster. As soon as an answer pops up in your head, you can click on the "Next" button. Once you click on the button, four answer options will be shown. You'll get 2.5 more seconds to select an answer from the list.

For the deliberative trials, the problem will be presented for up to 25 seconds so you can take the time to actively reflect on the problem. Whenever you feel you have reflected enough on the problem and decided on a final answer you can click on the "Next" button. Then four answer options will be shown and you will get 2.5 more seconds to select an answer from the list.

The answer options are only shown briefly to make sure you already calculated a response before. After you make your choice and click on it, you will be automatically taken to the next problem.

We'll clarify things further with a couple of practice problems. Press "Next" if you are ready to start the practice session!

These instructions were followed by two "passive viewing" trials, one fast and the other slow, where participants watched a full trial sequence (they could not click to advance) to familiarize them with the presentation times and the response deadline. This was followed by two active fast trials and two active slow trials where they had to respond.

At the end of practice, we clarified that the trials were grouped in blocks of 12 trials and that participants would be informed about the type of block (i.e., "fast" intuitive or "slow" deliberative) before each block, and would get a break after each block. To visually remind participants of which block they were solving, instructions and trials in an intuitive block were presented in green font, and in blue font in deliberate blocks. Finally, they were informed about the respective order of the blocks according to the speed condition (i.e., slow condition: "You'll get to do the deliberative blocks first, followed by the intuitive blocks."; fast condition: "You'll get to do the intuitive blocks first, followed by the deliberative blocks."; mixed condition: "The order will alternate between intuitive and deliberative blocks"). The first block then followed.

Every trial started with a fixation cross shown for 1 second. We then presented the first sentence of the bat-and-ball problem for 2 seconds. The second line (second sentence and question) was then displayed under the first sentence for 25 seconds (slow trial) or 4 seconds (fast trial) before the response options were displayed alone. Participants could also click on the Next button before the end of the presentation time to display the response options faster. Once those were displayed, participants had 2.5 seconds to select one of them by clicking on it. In case they failed to respond within the time limit, a message reminded them to answer before the deadline.

After the 12th trial of a block was completed, a message informed participants of their progress before they could proceed further (e.g., "You have completed a set of 12 deliberative trials, thanks! Press "Next" if you are ready to continue."). An additional message informed them when they were halfway through (i.e., "You finished 5 blocks out of 10, you are halfway through. Relax for a second and try to stay focused for the rest of the study.").

After participants had completed all 10 blocks, they were shown the original bat-and-ball problem and were asked whether they had seen it before. We also asked them to enter the solution. At the end of the study, participants completed a page with demographic questions.

Note that for piloting purposes, right before the demographics questions we also explained the bat-and-ball problem to participants and asked them to solve two more problems. These data were not analyzed for the current study.

1.4 Exclusion criteria

Missed deadlines. We discarded 285 trials (1.9% of 14700) where participants failed to select a response before the deadline. On average each participant contributed 24.4 (SD = 1) fast conflict trials, 24.3 (SD = 1.7) slow conflict trials, 24.3 (SD = 0.7) fast no-conflict trials, 24.8 (SD = 0.4) slow no-conflict trials, 9.7 (SD = 0.7) fast filler trials, and 9.7 (SD = 0.6) slow filler trials.

Bat-and-ball familiarity. The bat-and-ball is widely used and has been popularized in the media as well (Hoover & Healy, 2017). If participants already knew the task, they might not show any difference between slow and fast trials, which would dilute a potential boost in fast trial performance in the slow first condition. We therefore further excluded from our analysis 2242 trials out of 14415 (15.6%) from 19 participants who reported having seen the original problem before and were able to provide the correct "5 cents" response at the end of the experiment.

Fastest trials. Lastly, as suggested by one reviewer, we also removed 7 trials with reasoning or selection RT faster than 400 ms (about 2 SDs below the mean) to discard trials that possibly resulted from misinput or inattentiveness.

2 Results and discussion

Data was processed and analyzedusing the R software (R Core Team, 2017) and the following packages (in alphabetical order): *dplyr* (Wickham, François, Henry, & Müller, 2020), *ez* (Lawrence, 2016), *ggplot2*, (Wickham, 2016), *Rmisc* (Hope, 2013), and *tidyr* (Wickham & Henry, 2020).

2.1 Accuracy

We ran a 2 (type of trial, fast or slow) x 3 (order condition: fast first, slow first, mixed) ANOVA on the critical conflict trial accuracy with type of trial as within-subject and order condition as between-subjects factor. Results showed that the main effect of the order condition was not significant, $F(2,101) = 1.135, p = .326, \eta_p^2 = .02$, but the main effect of type was, $F(1,101) = 5.681, p = .019, \eta_p^2 = .05$. The interaction between the order condition and type was also significant, $F(2,101) = 7.202, p = .001, \eta_p^2 = .12$.

As Figure 1 shows, the main effect of trial type indicates that participants were overall slightly more accurate on the slow (M = 23.5%, SD = 40.7%) than on the fast (M = 19.9%, SD = 36.2%) trials. The interaction between trial type and condition indicates that this tendency was most pronounced in the fast first condition. Simple effect tests[3] indicated that the effect of trial type was significant in the fast first condition, $F(1, 31) = 7.144$, $p = .012$, $\eta_p^2 = .19$, but not for the slow first condition, $F(1, 36) = 1.529$, $p = .224$, $\eta_p^2 = .04$, nor the mixed condition, $F(1, 34) = 1.778$, $p = .191$, $\eta_p^2 = .05$. However, even in the fast first condition the effect was small with a 12.7% improvement in the slow trials. This finding fits with observations with the classic two-response paradigm and indicates that giving people the opportunity to deliberate hardly changes their accuracy. That is, by and large participants who manage to solve the problem correctly when they deliberate also manage to solve the problem when they have to reason fast. Participants who are biased when they have to intuit, typically stay biased when they are allowed to deliberate. This observation is further supported by our individual level analysis (see further).

However, our main interest concerned the order effect. Given that there was no main effect of the order factor and an order x type interaction we ran separate univariate ANOVAs on the order factor for the slow and fast trials. Our main a priori hypothesis concerned a possible order effect on the fast trials. Does solving slow trials before the fast trials leads to a performance boost on the fast trials? Figure 1 (left panel) shows the results. As the figure indicates, the differences between the conditions were small. A univariate ANOVA with the condition factor (fast first, slow first, or mixed) confirmed that the presentation order did not significantly affect the accuracies on the fast conflict trials, $F(2, 101) = 0.392, p = .677, \eta_p^2 = .01$). Pairwise Tukey comparisons also indicated that none of the contrasts between conditions reached significance, all $p > .6$. As Figure 1 suggests, if anything solving the slow trials first actually resulted in a slightly lower accuracy on the subsequent fast trials than when reasoners started with the fast trials (or when fast and slow blocks were alternated in the mixed condition). This directly argues

[3] As suggested by one of the reviewers, for completeness, we have followed-up all our significant interactions tests with simple effect tests.

against the hypothesis that the possibility for repeated deliberation in the slow trials will artificially boost the intuitive reasoning performance in the fast trials.

Although we had no a priori hypotheses with respect to an order impact on the slow trials, for completeness, we also explored the impact of the order factor on the accuracy on slow trials. Figure 1 (right panel) shows the results. As the figure suggests, there was a trend towards a better performance on the slow trials when the fast trials were solved first. However, a univariate ANOVA with the condition factor (fast first, slow first, or mixed) on the accuracies on the slow conflict trials indicated that the trend was not significant, $F(2, 101) = 2.275, p = .108, \eta_p^2 = .04$. Pairwise Tukey comparisons also indicated that none of the contrasts between conditions reached significance, all $p > .08$.

Finally, note that accuracies on the control no-conflict trials were very high (overall fast accuracy: $M = 98.1\%, SD = 4.4\%$; slow accuracy: $M = 98.9\%, SD = 4\%$, see supplementary Figure S2 for a full overview). For completeness, we also ran a 2 (trial type) x 3 (order condition) ANOVA on the no-conflict accuracies. The main effect of the order condition was not significant, $F(2, 101) = 1.307, p = .275, \eta_p^2 = .03$, but the main effect of type was, $F(1, 101) = 6.446, p = .01, \eta_p^2 = .06$ (i.e., overall 0.8% higher accuracy on slow trials). The interaction between the order condition and type was not significant, $F(2, 101) = 0.986, p = .377, \eta_p^2 = .02$. Note that the very high performance on the no-conflict control confirms that participants were paying minimal attention to the task and refrained from mere random responding (which would result in an average accuracy rate of 25%).

Nevertheless, one might note that the average accuracy on conflict trials hovered around 25%. Hence, in theory one could argue that participants might have specifically guessed when facing the conflict trials only. However, if participants guessed then our four response options (correct, heuristic, and two foils) should have been selected with equal frequency. More specifically, all our incorrect responses should be randomly distributed across the heuristic and two foil options. However, this was clearly not the case. Incorrect conflict responses were almost exclusively of the "heuristic" type (97.5% overall, 96.9% for fast trials and 98.3% for slow trials). Moreover, we also computed a stability index (e.g., Bago & De Neys, 2017). For each participant we calculated how frequently htey gave the same kind of response (i.e., heuristic, correct, or either of the two foils). Results showed that on average, participants gave the same type of response on 98.3% of conflict trials. Hence, individuals were highly robust in their choices. Taken together these results

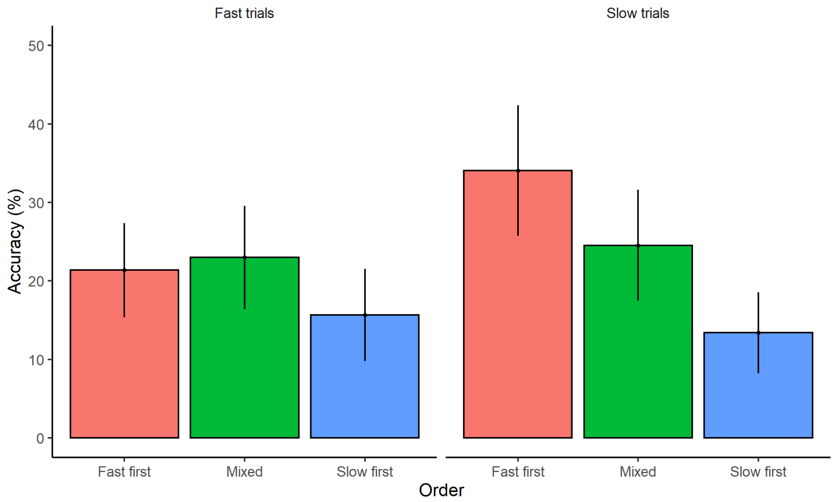

Figure 1. Average conflict accuracy by order condition, separated by type of Trial. Error bars show standard errors.

clearly argue against a guessing confound.

2.2 Latencies.

Reasoning latencies. In our fast trials, participants were instructed to respond as fast as possible, whereas they were allowed to reflect on their answer (up to 25 s) in the slow trials. We wanted to explore whether participants actually reasoned longer in the slow than in the fast conflict trials.[4] All latencies were log transformed prior to statistical analysis. Figure 2 (top panel) reports average back-transformed values for ease of interpretation. As with the accuracy analysis, we ran a similar 2 (type) x 3 (order) ANOVA on the reasoning latencies. Results showed there was indeed a main effect of trial type, $F(1, 101) = 80.899, p < .001, \eta_p^2 = .44$. There was no main effect of the condition factor, $F(2, 101) = 2.662, p = .075, \eta_p^2 = .05$, but there was a condition by type interaction, $F(2, 101) = 13.785, p < .001, \eta_p^2 = .21$.

The main effect of trial type indicates that people responded faster in the fast ($M =$ 2055 ms, $SD = 812$ ms) than in the slow ($M = 3216$ ms, $SD = 3057$ ms) conflict trials. The

[4]As with the accuracies, the interested reader can find an overview of response latencies on no-conflict problems in the Supplementary material.

interaction between trial type and condition indicates that this was less pronounced in the different conditions. In decreasing order, participants responded faster on fast trials than on slow trials by 1710 ms in the mixed condition, by 1274 ms in the slow first condition, and by 431 ms in the fast first condition. Simple effect tests of the interaction indicated that the effect of trial type was not significant in the fast first condition, $F(1, 31) = 1.281$, $p = .266$, $\eta_p^2 = .04$, but it was significant for the slow first condition, $F(1, 36) = 143.836$, p ¡ .001, $\eta_p^2 = .80$, and the mixed condition, $F(1, 34) = 25.119$, p ¡ .001, $\eta_p^2 = .42$.

One might note that although participants tended to reason longer on our slow trials, participants did not take much longer (i.e., overall +1161 ms, even in the slow first condition only 1274 ms) when given the time to reflect. This results is consistent with traditional two-response paradigm findings that already indicated that participants typically spend little time rethinking their answer when given the opportunity to do so in the final, deliberate response stage (e.g., Thompson et al., 2011; Bago & De Neys, 2017; Raoelison & De Neys, 2019). That is, alloting people more time to reflect per se does obviously not imply they will necessarily use much of this extra time. Reasoners can prefer to stick to fast, intuitive processing even when they are allowed to deliberate.

Response selection latencies. The above latencies concerned participants' reasoning times (i.e., the time elapsed between presentation of the question and participants' mouse click that allowed them to enter a response). To minimize backward inferencing, in the current design the actual response options were only briefly presented (i.e., 2.5 s) after the reasoning stage. In addition to the reasoning latencies, we also recorded participants' actual response selection latencies (i.e., time elapsed between presentation of the response options and mouse click on the selected option) to help rule out a potential complication. In theory, the selection phase might introduce a confound in the fast trials. When faced with reasoning time restrictions in the fast trials, participants might use the selection time for additional deliberation and thereby potentially boost their fast trial performance. However, if this were the case, we would expect considerably longer response selection latencies in the fast than in the slow trials. As with the reasoning latencies, we ran a 2 (type) x 3 (order) ANOVA on conflict selection latencies. Figure 2 (bottom panel) shows the results. There was no main effect of order condition, nor trial type, both F ¡ 1. As with the reasoning latencies, there was a type by order condition interaction, $F(2, 101) = 13.251, p < .001, \eta_p^2 = .21$.Simple effect tests of this interaction

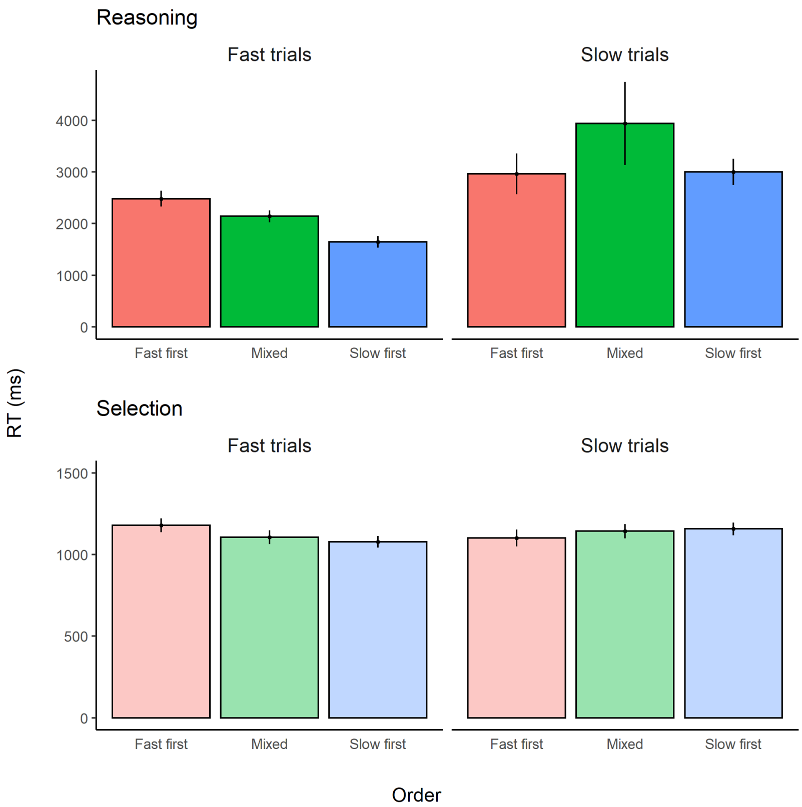

Figure 2. Overview of average reasoning and selection conflict reaction times (RT). Error bars show standard errors. *Note.* Time scales for reasoning (top) and selection (bottom) RTs are different for readability.

indicated that the effect of trial type was significant in the fast first condition, $F(1, 31) = 9.839$, $p = .004$, $\eta_p^2 = .24$, in the slow first condition, $F(1, 36) = 9.714$, $p = .004$, $\eta_p^2 = .21$, and in the mixed condition as well, $F(1, 34) = 4.482$, $p = .042$, $\eta_p^2 = .12$. As Figure 2 suggests, participants in the fast first condition were slightly slower (76 ms) in fast than slow trials, whereas the slow first condition (59 ms) and mixed condition (31 ms) showed the opposite effect. That is, except for the fast first condition, participants did not take longer to select their response in the fast than in the slow trials. But obviously, even in the fast first condition it is highly unlikely that a mere 76 ms longer selection time sufficed for any proper additional deliberation.

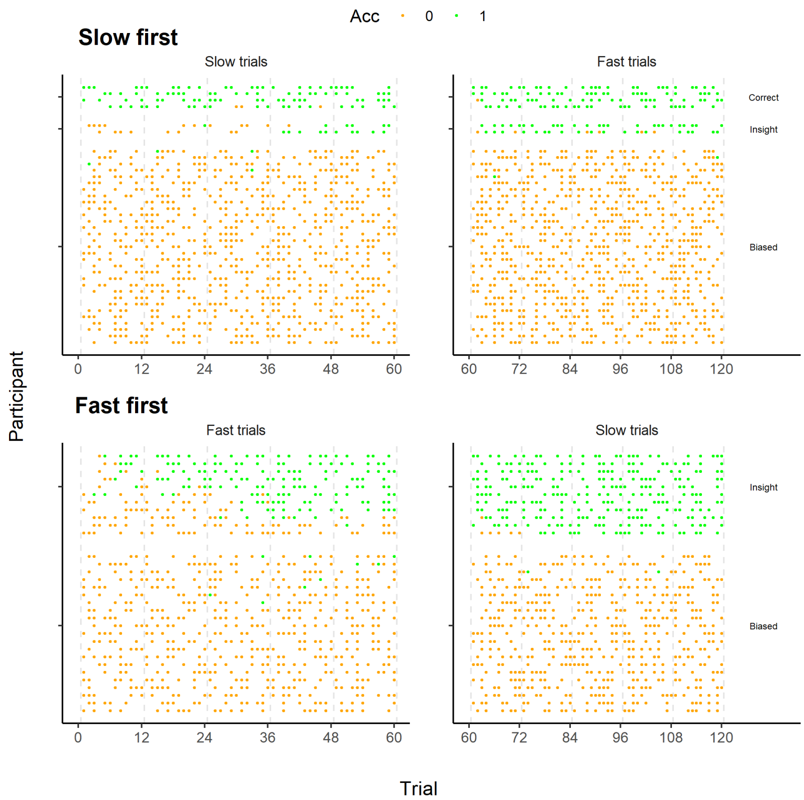

Figure 3. Individual trajectory for conflict items in the slow first and the fast first conditions. Labels indicate the group each participant belongs to. Vertical dashed lines separate blocks. *Note.* There were 4 correct, 2 insight, and 31 biased participants in the slow first condition; there were 11 insight and 21 biased participants in the fast first condition.

2.3 Individual level accuracy analysis

To explore further how participants solved the problems, we additionally performed an individual level accuracy analysis for each individual participant on each individual conflict trial from the start to the end of the experiment. This allows us to see how each individual participant's responding changed (or not) throughout the experiment. Figure 3 presents the results for the slow first and fast first conditions. As the figure indicates, by and large, we can classify the participants in three groups. First, most participants predominantly gave incorrect responses from start to finish. Hence, the majority of the participants consistently gave incorrect fast and slow responses and remained biased throughout the study. This group is labeled as the "biased" group in Figure 3. Second, a smaller group of participants gave a correct response at their very first trial (fast or slow) and predominantly remained responding correctly throughout the study. This group of "correct" reasoners did not need any learning to arrive at the correct answer. Third, there is a small number of participants that started with an incorrect response and found the correct answer somewhere along the way. Once they had found the solution, they remained correct on almost all subsequent trials. We labeled this group as the "insight" group. What is critical is that this group was especially small in the slow first condition (i.e., $n = 2$, 5% of total). This validates our overall analysis and indicates that there is little support for the claim that repeated deliberation (i.e., in the slow trials) subsequently boosts intuitive performance in the fast trials. Indeed, spontaneous insight with repeated presentations is even more likely in the fast first group (i.e., $n = 11$, 34% of total). Hence, a small group of reasoners can show improved performance with repeated presentation but this does not necessarily result from slow deliberation. This insight can also occur when people repeatedly intuit.[5] However, the bottom-line is that whether or not people are allowed to deliberate, spontaneous insight remains exceptional. The vast majority of participants respond exactly the same (biased or correct) from start to finish.

Note that for illustrative purposes Figure 3 only presents the slow first and fast first conditions. A separate graph including the mixed condition can be found in the Supplementary material(Figure S4). Results for the mixed group are fully consistent with the other two conditions.

General discussion

In this study we manipulated the order of presentation of fast and slow trials to test for a potential confound in the classic two-response paradigm (Thompson et al., 2011), where participants have to give a fast, intuitive answer immediately followed by a slow, deliber-

[5]During the fast blocks, participants solved one problem immediately after the other. However, one could argue that participants had time to deliberate during the breaks between blocks. In theory, insight for people in the fast first condition might have occurred during such "break deliberation". Note, however, that as Figure 3 indicates, contrary to this account, the correct responding in the insight group typically started midway through the blocks and not immediately after the breaks.

ate one. The fact that people are repeatedly asked to deliberate on multiple trials, might allow them to learn and automatize the correct solution strategy. This repeated deliberation would thereby artificially boost people's intuitive reasoning performance. Therefore, Markovits et al. (2019) presented an alternative "two-block" paradigm in which participants were first presented with a full block of "fast" intuitive trials and only afterwards they were given a block of "slow" deliberate trials in which they could take more time to solve the problem. Here we tested directly whether allowing people to repeatedly deliberate will boost their intuitive reasoning performance by manipulating the order of the fast and slow blocks. Results clearly indicated that repeatedly having people deliberate first did not boost their accuracy on the fast, intuitive trials.

These results lend credence to previous findings in the literature that adopted the two-response paradigm and indirectly argued against the repeated intermittent deliberation confound (e.g., Bago & De Neys, 2017, 2019b; Lawson, Larrick, & Soll, 2020; Raoelison & De Neys, 2019; Thompson & Johnson, 2014; Thompson et al., 2011). For example, Bago and De Neys (2019b) noted that a control analysis that was restricted to the first study trial led to the same pattern of results as the overall data, suggesting that the deliberation stage of the first trial(s) did not boost performance in later trials. Similarly, Raoelison and De Neys (2019) found that after very extensive repeated presentation of the bat-and-ball problem even accuracy in the final, deliberate two-response stage rarely improved (See also Hoover & Healy, 2017, for related findings).

Taken together, this argues against the suggestion that all intuitive trials need to be presented first before allowing participants to deliberate. Allowing people to deliberate before they are asked to intuit does not artificially boost people's intuitive reasoning performance. On the contrary, if anything, we observed the opposite trend for slow trials (better performance on slow trials after solving fast trials first). Although speculative, this may indicate that asking participants to intuit before allowing them to deliberate might actually increase deliberate performance. Our individual level accuracy analysis also revealed that more insight happened when fast trials were presented first. However, the trends were only suggestive and our study was not specifically designed to test this hypothesis. The possible trend should therefore be interpreted with caution. At the same time, the trend should not be discarded either. Our results argue against the claim that repeated deliberation boosts intuitive reasoning performance. In this respect the results help to validate the two-response paradigm. However, the results do not imply that asking people to both provide an intuitive and deliberate response intra-trial does not affect reasoning per se. Although it was unexpected and we have no clear theoretical explanation, repeated intuiting might actually help to deliberate better. In this sense, the two-response paradigm would over- (rather than under) estimate the contribution of deliberation in the reasoning process (i.e., as measured with a traditional "one-response" paradigm). People might deliberate even less when they are not explicitly instructed to provide an intuitive response first. However, as we noted, this hypothesis would need to be tested further. Our claims in this paper primarily concern the impact of repeated deliberation on people's intuitive reasoning performance and not vice versa.

One might note that both in the current blocked design and the original two-response paradigm, participants are–by definition–asked to repeatedly switch from one task ("reason intuitively") to the other ("reason deliberately"). Such task switching might come with its own costs (e.g., Kiesel et al., 2010; Monsell, 2003). Previous work has suggested that overall "deliberate" accuracies in a traditional one-response design do not diverge from those observed in the two-response paradigm (e.g., Bago, Bonnefon, & De Neys, 2020; Bago & De Neys, 2019a; Thompson et al., 2011). This argues against a key confound due to repeated switching. However, the precise role of task switching per se remains to be explored in detail and this might open an interesting avenue for futures studies.

In closing, it should be clear that our current critique of Markovits et al. (2019) concerns the suggestion that the fast-trials-first design is needed to get a "pure" measure of intuitive reasoning. We do not contest that Markovits et al.'s suggestion to use separate fast and slow trials is a useful methodological alternative to the traditional two-response paradigm in which intuitive and deliberate response stages are manipulated intra-trial. This design can help validate two-response findings from the two-response paradigm (e.g., do we get similar results with different paradigms?) and offers practical advantages (e.g., much easier to implement a simple, pen-and-paper version in a classroom setting, clearer temporal separation of intuitive and deliberate processing in a neuroimaging setting). We readily acknowledge that the inter-trial manipulation and temporal segregation of intuitive and deliberate trials that Markovits et al.'s fast-slow paradigm provides, are a welcome addition to our collective toolbox. However, there is no good evidence for the claim that the paradigm is needed to prevent that repeated deliberation will artificially boost people's intuitive reasoning performance.

Acknowledgments

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

Open practices statement

Raw data can be downloaded from our OSF page (https://osf.io/pgk7t/).

References

Bago, B., Bonnefon, J.-F., & De Neys, W. (2020). Intuition rather than deliberation determines selfish and prosocial choices. *Journal of experimental psychology: General*. Advance online publication. doi:10.1037/xge0000968

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

Bago, B., & De Neys, W. (2019a). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of experimental psychology: General*, *148*(10), 1782–1801. doi:10.1037/xge0000533

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

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

De Neys, W. (Ed.). (2017). *Dual process theory 2.0*. doi:10.4324/9781315204550

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. doi:10.1177/0963721419855658

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

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

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. doi:10.1177/1745691612460685

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*(4), 25–42. doi:10.1257/089533005775196732

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. doi:10.3758/s13423-017-1241-8

Hope, R. M. (2013). *Rmisc: Rmisc: Ryan miscellaneous*. R package version 1.5. Retrieved from https://CRAN.R-project.org/package=Rmisc

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

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching : A review. *Psychological Bulletin*, *136*(5), 849–874. doi:10.1037/a0019842

Lawrence, M. A. (2016). *Ez: Easy analysis and visualization of factorial experiments*. R package version 4.4-0. Retrieved from https://CRAN.R-project.org/package=ez

Lawson, M. A., Larrick, R. P., & Soll, J. B. (2020). Comparing fast thinking and slow thinking: The relative benefits of interventions, individual differences, and inferential rules. *Judgment and Decision Making*, *15*(5), 660–684.

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

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. doi:10.3758/s13423-017-1234-7

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. doi:10.1016/S1364-6613(03)00028-7

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

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/

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

Rosas, A., & Aguilar-Pardo, D. (2019). Extreme time-pressure reveals utilitarian intuitions in sacrificial dilemmas. *Thinking & Reasoning*, *0*(0), 1–18. doi:10.1080/13546783.2019.1679665

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(2), 215–244. doi:10.1080/13546783.2013.869763

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

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. doi:10.1016/j.cogpsych.2011.06.001

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the cognitive reflection test. *Cognition*, *150*, 109–118. doi:10.1016/j.cognition.2016.01.015

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation*. R package version 0.8.5. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Henry, L. (2020). *Tidyr: Tidy messy data*. R package version 1.0.2. Retrieved from https://CRAN.R-project.org/package=tidyr

Supplementary material

A Counterbalancing and randomization

Figure S1 illustrates how we counterbalanced conflict content and the randomization/ordering process. Panel A illustrates the conflict content counterbalancing in the fast first condition: conflict (no-conflict, respectively) items from the first block had their corresponding no-conflict (conflict, respectively) versions in the sixth block, and so on. Panel B shows the two sets used in the slow first condition to avoid a potential confound between fast/slow condition and content. Each block had a fast and a slow version which were used in the two sets for all conditions. Panel C illustrates the ordering of blocks in each condition. In the fast first and the slow first conditions, the blocks within the fast or slow part were randomized.

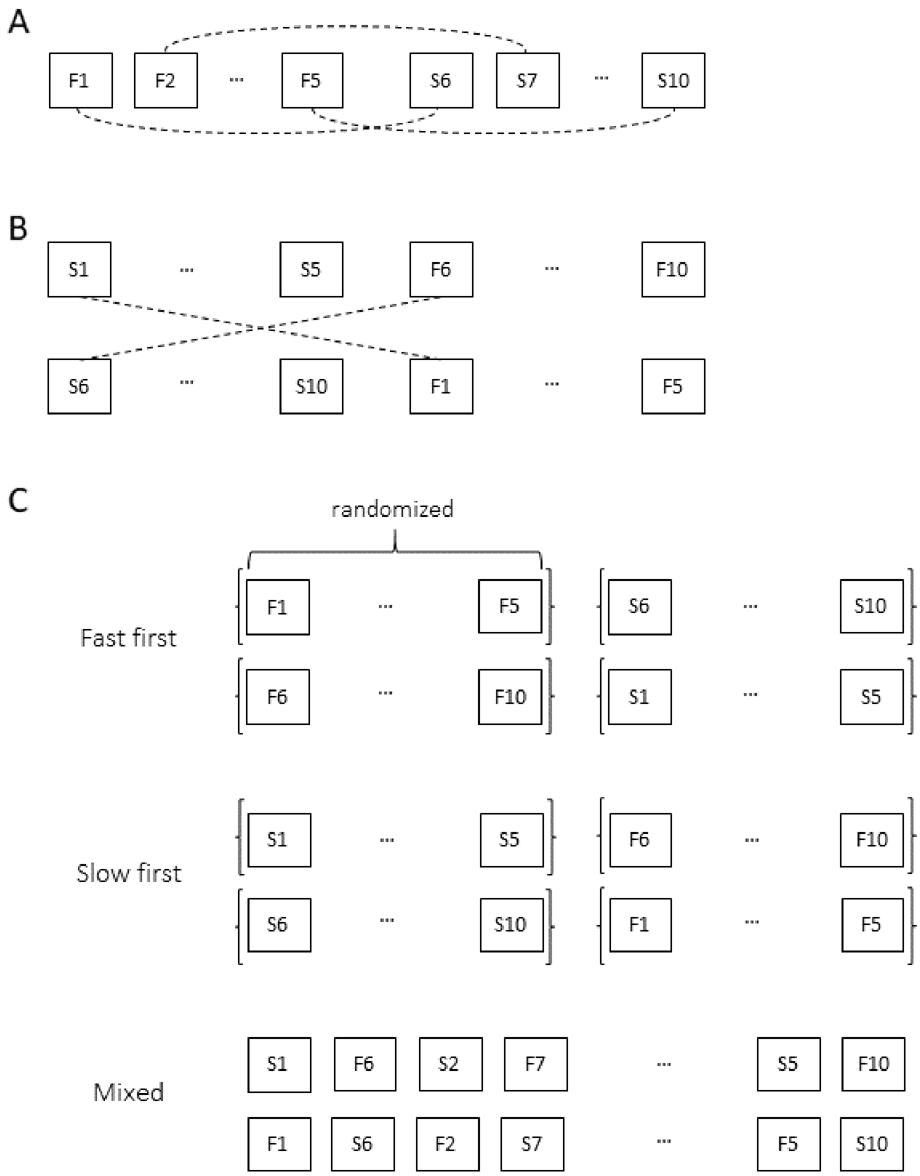

Figure S1. Block organization. Panel A: conflict counterbalancing; Panel B: fast/slow counterbalancing; Panel C: block ordering.

B Additional figures

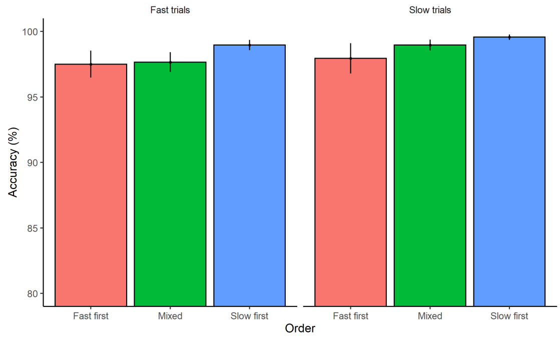

Figure S2. No-conflict accuracy for all order conditions, separated by type of trials. Error bars are standard errors.

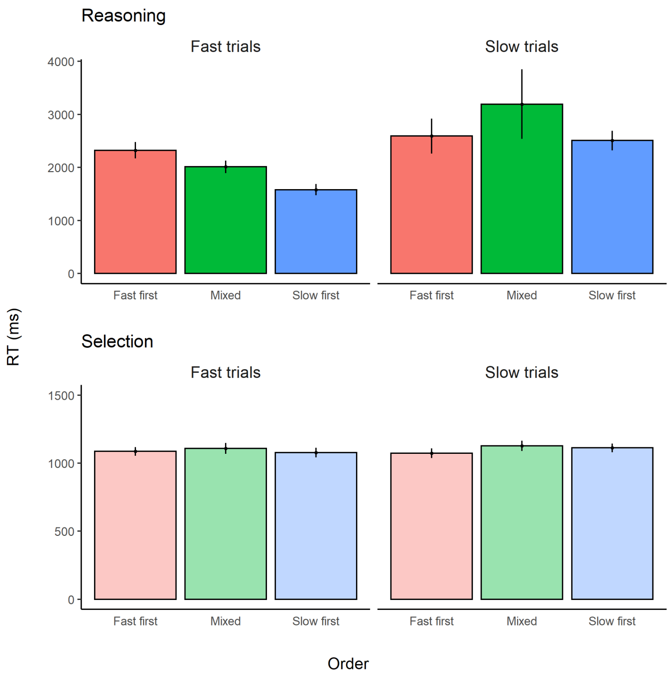

Figure S3. Average reasoning (top panel) and selection (bottom panel) no-conflict latency by order conditions, separated by type of trial. Error bars are standard errors. *Note.* Time scales are different for readability.

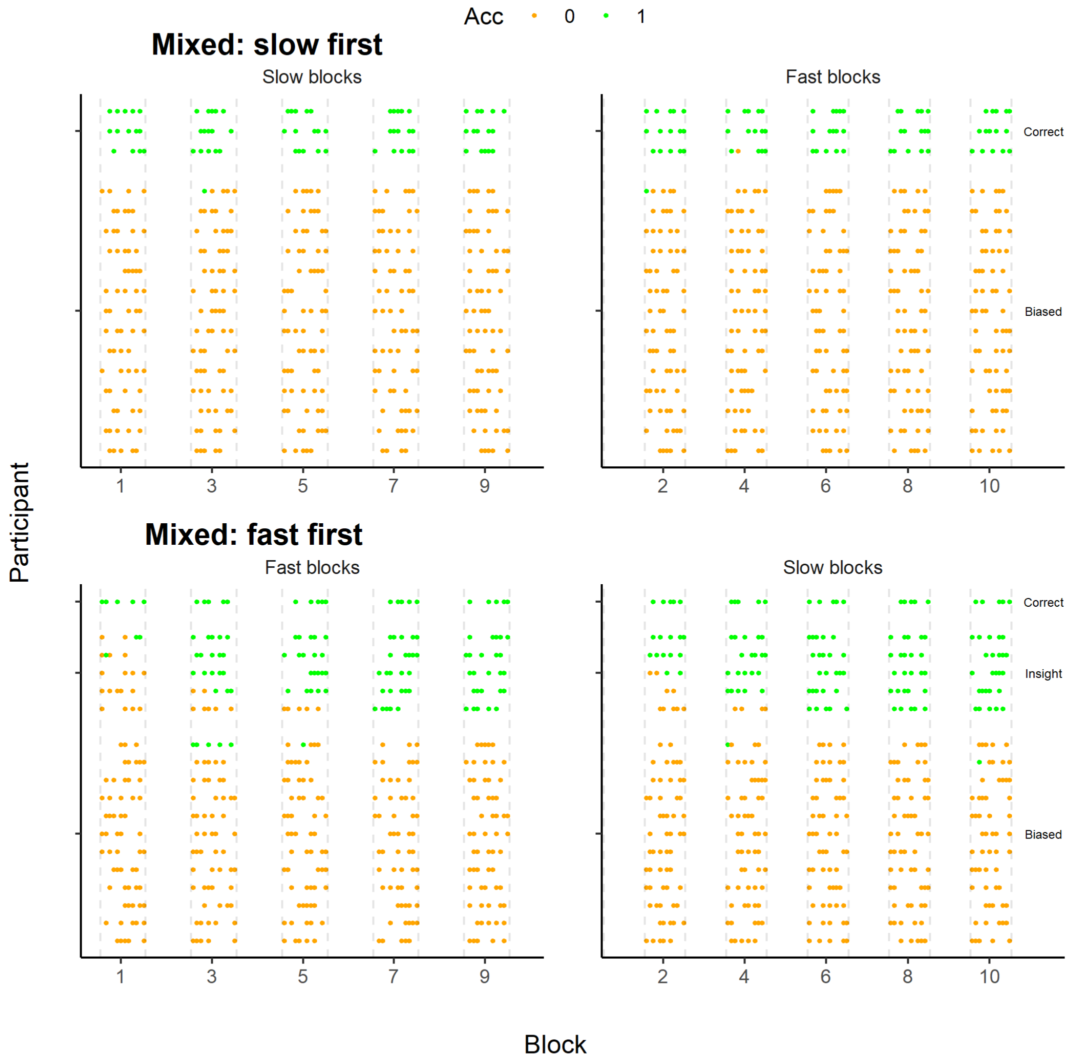

Figure S4. Individual trajectories for the mixed order condition, separated by type of block. Labels indicate the group to which each participant belongs. Vertical dashed lines separate blocks. The mixed condition has been separated in two groups according to the first type of block participants were presented with (mixed: slow first and mixed: fast first). *Note.* There were 3 correct and 14 biased participants in the mixed, slow first group; there were 1 correct, 5 insight, and 12 biased participants in the mixed, fast first group.