Conflict detection predicts the temporal stability of intuitive and deliberate reasoning

Aikaterini Voudouri, Michal Bialek, Artur Domurat, Marta Kowal, Wim de Neys

Conflict detection predicts the temporal stability of intuitive and deliberate reasoning

https://doi.org/10.1080/13546783.2022.2077439

Aikaterini Voudouri[1], Michał Białek[2], Artur Domurat[3], Marta Kowal[2], Wim De Neys[1]

[1] Université Paris Cité, LaPsyDÉ, CNRS, F-75005 Paris, France
[2] Institute of Psychology, University of Wrocław, Wrocław, Poland
[3] Center for Economic Psychology and Decision Sciences, Kozminski University, Poland

Abstract

Although the susceptibility to reasoning biases is often assumed to be a stable trait, the temporal stability of people's performance on popular heuristics-and-biases tasks has been rarely directly tested. The present study addressed this issue and examined a potential determinant for answer change. Participants solved the same set of "bias" tasks twice in two test sessions, two weeks apart. We used the two-response paradigm to test the stability of both initial (intuitive) and final (deliberate) responses. We hypothesized that participants who showed higher conflict detection in their initial intuitive responses at session 1 (as indexed by a relative confidence decrease compared to control problems), would be less stable in their responses between session 1 and 2. Results showed that performance on the reasoning tasks was highly, but not entirely, stable two weeks later. Notably, conflict detection in session 1 was significantly more pronounced in those cases that participants changed their answer between session 1 and 2 than when they did not change their answer between sessions. We discuss practical and theoretical implications.

Keywords: dual-process theory; conflict detection; two-response paradigm; heuristics-and-biases

Introduction

Although reasoning has been characterized as the essence of our being, it is often prone to cognitive biases. Decades of research in the reasoning and decision making fields have shown that when faced with simple reasoning tasks, people tend to overlook their underlying logical principles and, as a result, provide incorrect answers (Kahneman, 2011). Consider the following problem:

Imagine you are running a race. If you pass the person in second place what place are you in?

The answer that often pops into mind is "first place". However, if one takes the time to further reflect on the problem, it is clear that the correct answer is in fact "second place". Despite the simplicity of the solution, mistakes in reasoning tasks like the above are very frequent. This is because people often base their answer on mental shortcuts (e.g., "after second comes first" in the above example), instead of providing an answer that agrees with logical norms (e.g., "if you pass the second runner, there is still a person ahead of you"). A prevalent explanation as to why these errors of judgement happen, has been proposed by dual-process theories. These theories view reasoning as an interaction

between two systems, System 1 and System 2, which approximately correspond to intuitive and deliberate thinking (e.g., Epstein, 1994; Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996). The main difference between these systems is that while System 1 is autonomous and does not make use of cognitive resources, System 2 requires cognitive resources to operate. System 1 can be helpful in many cases (e.g., when a decision has to be taken quickly), but it also often cues "heuristic" answers, responses that are based on rules of thumb, stored associations, and stereotypes. Classic dual process theories support that when a problem cues a "heuristic" answer that conflicts with logical considerations, reasoners need to engage in effortful thinking and further contemplate the problem in order to override their "intuitive", erroneous answer and provide a normative response[1] (Evans & Stanovich, 2013; Kahneman, 2011). However, in most cases, in order to minimize effortful thinking reasoners stick to their "heuristic" answer and respond incorrectly (Evans & Over, 1996; Kahneman, 2011).

Heuristic biases have been widely researched in the literature and have been predicted using a range of cognitive tasks (Białek et al., 2020; Šrol & De Neys, 2021; Stupple et al., 2013; Toplak et al., 2014). Nevertheless, it is not completely clear whether the performance of reasoners on bias tasks is stable over time. Although bias susceptibility is generally assumed to be a stable individual trait, in the sense that biased reasoners are thought to remain biased from one moment in time to another, reasoners' response consistency has been rarely directly tested (e.g., Białek & Pennycook, 2018; Meyer et al., 2018; Stango & Zinman, 2020). In the present paper, we will investigate this consistency and discuss a potential determinant for answer change.

The determinant we will focus on is reasoners' detection of conflict between competing responses (e.g., De Neys & Glumicic, 2008; De Neys, 2012; Šrol & De Neys, 2021). Over the last decade, numerous studies have indicated that when people solve classic "bias" tasks in which they are faced with a cued heuristic response that conflicts with logical principles, they often show some sensitivity to this conflict (e.g., Bago & De Neys, 2017; De Neys, 2014; De Neys et al., 2013; Gangemi et al., 2015; Mata, 2020; Pennycook et al., 2015; Stupple et al., 2013; but see also Ferreira et al., 2016; Mata et al., 2017; Pennycook et al., 2012). For example, reasoners typically show lower confidence when answering a classic "bias" task than when solving a control version in which the cued heuristic does not conflict with logical principles (e.g., a no-conflict version of the introductory race problem might read "Imagine you are running a race. If you pass the person in first place, what place are you in?"). This suggest that people detect, to some extent, that there are conflicting responses at play.

In this study, we wanted to explore if conflict detection is related to how often people change their answers on classic bias tasks from one point in time to another. The general idea was that the more conflicted reasoners feel about an answer, the more likely they might be to change this answer at a future time. Evidence for this comes from the two-response paradigm, where participants are asked to provide two consecutive responses to a problem (Thompson et al., 2011). During the first (initial) response stage participants see the problem and are asked to give the very first answer that comes to mind. Then, during the second (final) response stage, they are presented with the problem again and are asked to reflect on it before providing their final answer. Because of the instruction differences, the initial response is thought to be provided predominantly through System 1 processing with minimal System 2 involvement, while the final response is thought to be given predominantly through deliberate, System 2 processing (Thompson et al., 2011). In an attempt to minimize System 2

[1] When we refer to the "logical", "normative", or "correct" response we are referring to the response that has traditionally been considered to be correct according to standard logic and probability theory.

engagement during the initial stage, recent studies ask participants to provide their first response under a strict deadline and a cognitive load (e.g., a parallel task taxing their cognitive resources). Since System 2 requires cognitive resources to operate, these constraints force participants to provide their answers intuitively during the initial stage (Bago & De Neys, 2017). Hence, the two-response paradigm allows us to directly compare intuitive and deliberate responses on the same problem.

Studies using this paradigm have shown that the higher the conflict detection at the initial response stage, the more likely participants' answers are to change in the final stage (Bago & De Neys, 2017, 2020; Thompson & Johnson, 2014). In other words, if reasoners feel more conflicted (i.e., less certain) about their initial response, they are more likely to change it after they are given the time to deliberate. This (un)certainty about the initial response is also being referred to as the "Feeling of Rightness" (FOR, Thompson et al., 2011). That is, the lower the feeling of rightness (i.e., the confidence) that reasoners show at the initial answer, the more likely it is for them to reconsider their answer in the final stage (Thompson et al., 2011).

Recent dual process models have presented a new conceptualization to account for the conflict detection and two-response findings (Bago & De Neys, 2017, 2019a; De Neys & Pennycook, 2019; Handley et al., 2011; Pennycook et al., 2015, Newman et al., 2017; see De Neys, 2017, for review). In essence, these models postulate that the "logical" response that has traditionally been considered to be cued by System 2, can also be cued by System 1. The main idea is that System 1 can not only give rise to "heuristic" intuitions, which cue responses that contradict logic, but also to "logical" intuitions, which cue responses that are in line with logical principles. The latter are believed to be based on an intuitive/automated understanding of probabilistic and mathematical rules. The most dominant "type" of intuition (i.e., heuristic or logical) will be the one to eventually prevail. Let's imagine that the two competing intuitions–"heuristic" and "logical"–have a large difference in their activation levels, with one's strength dominating over the other's. In that case, there will be little conflict experienced when generating an initial response and it will be unlikely that the reasoner engages in deliberation and changes their response. Instead, if the two types of intuitions have very similar activation levels, conflict will be maximal and it will be more likely that the reasoner will engage in deliberation to correct their initial response (Bago & De Neys, 2019a; De Neys & Pennycook, 2019; Pennycook et al., 2015; Trippas & Handley, 2017).

Our rationale in the present study was that the same mechanism that drives answer change from the initial to the final response in a single trial, might also drive answer change across a longer time window, for example at different test occasions. Our aim was to explore whether the conflict detection at the initial, intuitive response of a given test session, is related to the response change at a later re-test session (both at the intuitive and at the deliberate level). The reasoning behind this is similar to the one described above: the more dominant one intuition is compared to its competitor (e.g., say, one is strength "9 out of 10" and the other is strength "2 out of 10"), the less conflict is created, and the more likely it should be that it will keep dominating over the weaker intuition at a future test occasion. The more similar the two intuitions are in strength (e.g., one is strength "5 out of 10" and the other is strength "6 out of 10"), the higher the conflict that is created, and the more likely it is that potential random noise (e.g., 1 unit variability due to participants' concentration, level of tiredness etc.) will reverse the strength ordering and make the other intuition dominate, thus, leading to answer change.[2]

[2] Since the dominance of two intuitions of similar strength can be reversed by random noise, we should note that this reversal can go both ways. More specifically, a participant's heuristic response at the first

Above we sketched the theoretical background that inspired our rationale. However, we can clarify the core idea with a simple non-theoretical analogy. Imagine one has a choice between two desserts; ice-cream or cupcakes. Person A really likes cupcakes, but dislikes ice-cream, while Person B likes both equally well. When you ask Person A about their decision, they will have little doubt about it given their dominant preference and, if you ask them again next week, it is very likely that they will make the same decision. Person B, however, will presumably face a hard decision since they like both desserts but they have to choose one. Whatever the final choice of Person B is, they will presumably be less confident that they made the right decision and it is more likely that they will choose differently if they are asked at another time in the future. It is in this sense that we expect response conflict (or inversely response confidence) to be predictive of response stability. The stronger the preference, the less conflict or doubt there will be about the decision, so the more likely it is that one's choices will remain stable over time.

To test whether conflict detection can be predictive of response stability, we asked participants to solve a set of heuristics-and-biases tasks (test session 1), and re-contacted them again two weeks later to solve the same tasks again (test session 2). We used the two-response paradigm for both test sessions. We hypothesized that participants who showed higher conflict detection in their initial, intuitive response at session 1, would be less stable in their responses between session 1 and session 2 (both at the intuitive and the deliberate level). For the calculation of conflict detection we focused on initial trials, as they offer a purer measure of conflict that is independent of deliberation (Bago & De Neys, 2019a).

Method

Preregistration

The study design and research question were preregistered on the Open Science Framework (https://doi.org/10.17605/OSF.IO/8FN3U). No specific analyses were preregistered.

Participants

We recruited our participants online on Prolific Academic (www.prolific. ac). 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. There were two test sessions that were two weeks apart. Participants were re-contacted two weeks after the first test session. The second session was not announced during session 1. Hence, participants were not aware that they were going to be re-tested before they were re-contacted. Participants were paid respectively £1.7 and £2 for their participation in session 1 and 2.

We initially recruited 200 participants of which 132 completed both test sessions. Of these 132, 60 had to be discarded because of a randomization coding error. We therefore recruited an additional 100 participants of which 79 completed both test sessions. This resulted in a total sample of 151 participants who completed both test sessions as intended. The mean age of these participants was 36.5 years (SD = 13.4) and 60.2% of them were female. Thirty-eight percent had a high school degree as their

test session can be turned into a logical response at the re-test session and vice versa (i.e., a logical response at the test session can become a heuristic response at the re-test session).

highest education level and 47% had a bachelor's degree. All reported data concern the results of these 151 participants who completed both test sessions.

Materials

Counterbalancing. Participants were presented with four different reasoning tasks (i.e., bat-and-ball, base-rates, syllogisms and conjunction fallacy tasks). Each task was composed of eight conflict and eight no-conflict problems. For every reasoning task two sets of items were created in which the conflict status of each item was counterbalanced. More specifically, all the conflict items of the first set appeared in their no-conflict version in the second set, and all the no-conflict items in the first set appeared in their conflict version in the second set. Half of the participants were presented with the first set of problems while the other half was presented with the second set. This way, the same content was never presented more than once to a participant and everyone was exposed to the same items, which minimized the possibility that mere item differences influence the results. The presentation order of the tasks and the items within each task was randomized.

Bat-and-ball problems (BB). Each participant was presented with eight bat-and-ball problems in multiple-choice format (four conflict and four no-conflict) taken from Raoelison and De Neys (2019). Although the amounts and the names of the objects varied between items, all items shared the same structure with the classic bat-and-ball problem. Participants were always provided with two answer options; a logical answer ("5 cents" in the original bat-and-ball), which was also considered as correct, and a heuristic answer ("10 cents" in the original bat-and-ball), which was considered as incorrect. An example of the problems is presented below:

A national park has 650 roses and lotus flowers in total.
There are 600 more roses than lotus flowers.
How many lotus flowers are there ?

- *25*
- *50*

The no-conflict versions were constructed by removing the "more than" statement from the conflict versions. For instance, in its no-conflict version the above example would become "A national park has 650 roses and lotus flowers in total. There are 600 roses. How many lotus flowers are there?". Each problem was presented in two stages. First, the first sentence was presented for 2000 ms. Afterward, the second sentence along with the question and the answer options was added until a response was given or until the deadline. As in Bago and De Neys (2019), the deadline for the initial response was 4000 ms.

Base-rate problems (BR). The base-rates problem presentation format was based on Pennycook et al's (2014) rapid-response paradigm. The sentences of each problem were presented serially and the amount of text that was presented on the screen was minimized. Participants were presented with eight base-rate problems (four conflict and four no-conflict) taken from Pennycook et al. (2014). Each problem consisted of a sentence describing the composition of a sample (e.g., "This study contains businessmen and firemen."), a sentence with a stereotypical description of a random person from the sample (e.g., "Person 'K' is brave.") and a sentence with the base-rate information (e.g., "There are 996 businessmen and 4 firemen."). Participants were then asked to choose the group that the random person most likely

belonged to. The answer option that was considered correct was always the one that corresponded to the vast majority of the people in the sample. An example of the problems is presented below:

This study contains businessmen and firemen.
Person 'K' is brave.
There are 996 businessmen and 4 firemen.
Is Person 'K' more likely to be:

- *A businessman*
- *A fireman*

The no-conflict versions were constructed by reversing the base-rates of the conflict versions. For example in its no-conflict version, the second sentence of the above problem would read "There are 4 businessmen and 996 firemen". Each problem was presented in three stages. First, the first sentence was presented for 2000 ms. Then, the second sentence was added for another 2000 ms, and finally the critical base-rate information along with the question and the answer options were added until a response or until the deadline. As in Bago and De Neys (2017), the deadline for the initial response was 3000 ms.

Syllogistic reasoning problems (SYL). Each participant was presented with eight syllogistic reasoning problems, four conflict and four no-conflict, taken from Bago and De Neys (2017). Each problem consisted of a major premise (e.g., "All fruits can be eaten."), a minor premise (e.g., "Strawberries are fruits.") and a conclusion (e.g., "Strawberries can be eaten."). Participants were told to always consider the premises as true and were asked to say if the conclusion followed logically from the premises or not. A conclusion was considered logical only when it was valid. An example of the problems is presented below:

All fruits can be eaten.
Strawberries can be eaten.
Strawberries are fruits.
Does the conclusion follow logically?

- *Yes*
- *No*

In the conflict problems, the believability and the validity of the problems were in conflict, meaning that a syllogism was either valid and unbelievable or invalid and believable. For instance, in the above conflict problem the syllogism is believable, but invalid. On the contrary, in the no-conflict problems, the syllogisms were either valid and believable or invalid and unbelievable. For example, the valid and believable no-conflict version of the above problem would read: "All fruits can be eaten. Strawberries are fruits. Strawberries can be eaten". Each problem was presented in three stages. First, the first sentence of the problem was presented for 2000 ms. Then, the second sentence was added for 2000 ms., and finally the conclusion along with the question and the answer options were added until a response was given or until the deadline. As in Bago and De Neys (2017), the deadline for the initial response was 3000 ms.

Conjunction fallacy problems (CONJ). Each participant was presented with eight conjunction fallacy problems, four conflict and four no-conflict, that were taken from Frey et al. (2018), apart from one item (i.e., the Linda problem) which was adapted from the material of Tversky and Kahneman (1983). Each problem consisted of a stereotypical description of an individual followed by two statements about this individual, and participants were asked to choose the statement that was more

likely to be true. The first answer option consisted of a single statement related to the individual (e.g., "Jon plays in a rock band"), while the second response option was a conjunction of the first statement with a second statement (e.g., "Jon plays in a rock band and is an accountant"). One of the two statements had a strong fit with the stereotypical description, while the second one had a lower fit. Since the possibility of a single event occurring is always higher than the possibility of the conjunction, the single statement was always considered as the correct choice. An example of the problems is presented below:

John is 32.
He is intelligent and punctual but unimaginative and somewhat lifeless.
In school he was strong in mathematics but weak in languages and art.
Which statement is most likely:

- *John plays in a rock band*
- *John plays in a rock band and is an accountant*

The no-conflict versions were created by replacing the singular option with the statement that showed a strong stereotypical fit to the description. For instance, in the no-conflict version of the above example the two answer options would be : Option 1: "John is an accountant", Option 2: "John is an accountant and plays in a rock band". Each problem was presented in two stages. First, the first part of the problem (description) was presented for 4000 ms. Then the critical question and answer options were added and remained on screen until a response was given or until the deadline. The deadline for the initial response was 5000 ms (see Boissin et al., 2021).

Two-response format. We used the two-response paradigm (Thompson et al., 2011) for the presentation of all items. In this paradigm participants are asked to provide two consecutive responses on every trial (see Procedure). The paradigm's format was based on recent studies in which, during the initial response, participants are asked to perform a load memorization task as well as to respond under a strict deadline, which is pre-tested to be demanding for the respective task (e.g., Bago & De Neys, 2017, 2019a; Boissin et al., 2021; Raoelison et al., 2020). During the final response there is no load or deadline. As already mentioned, System 2 requires cognitive resources to operate, so by restricting the processing time and adding a memorization load during the first stage, System 2 involvement is minimized. As a result, one can be maximally sure that the initial response is provided intuitively (i.e., without deliberation), while in the final response stage reasoners are allowed to deliberate. The load memorization task that we used was a complex visual pattern (i.e., 4 crosses in a 3 × 3 grid) and it was briefly presented before each problem (Miyake et al., 2001). After providing an initial response, participants were presented with four different load patterns (i.e., with different cross placings) and had to identify the one that they had been asked to memorize.

Procedure

The experiment was run online using the Qualtrics platform. Participants were told that the study would take 20 minutes to complete and that it demanded their full attention. They were first presented with a general description of the task, where they were informed that they would have to provide two consecutive responses to various reasoning problems. More specifically, they were told to first answer with the very first answer that came to their mind and then reflect on the problem before providing their final response (see Bago & De Neys, 2017 for literal instructions). In order to familiarize themselves with the two response procedure, they first solved two simple mathematical problems (addition and subtraction) with the two response format. Then, they practiced the load task alone, by

solving two memorization trials. Finally, they practiced the two math problems in their full two-response format (problem + deadline and load task on initial response). After the practice, participants started the main task which consisted of four blocks and 32 reasoning problems (eight problems per block). Each block consisted of a single task (i.e., either bat-and-ball, base-rates, syllogisms or conjunction fallacies). At the start of each block participants received specific instructions for the respective task, they were shown an example problem and solved a practice problem. Each trial started with a fixation cross shown for 2000 ms. Then the first part of the problem was presented (for more details see Materials subsections for each reasoning task), followed by the matrix for the cognitive load task which remained on screen for 2000 ms. Then the whole problem was presented, along with the question and the answer options. Participants could provide their initial response by clicking on one of the answer options. One second before the deadline, the screen turned yellow to remind participants that the deadline was approaching. If they did not respond within the deadline, they were presented with a message asking them to try and respond within the deadline on the next trials. If they responded within the deadline, they were asked to rate their confidence in the correctness of their initial response on a scale from 0 (absolutely not confident) to 100 (absolutely confident). Immediately after, participants were shown four matrices and were asked to recall the test matrix. They were then given feedback on the correctness of their recall. Finally, participants viewed the full problem again and were asked to provide their final answer. Next, they were asked their confidence in the correctness of their final response.

Participants were re-contacted after two weeks to complete session 2 of the study, which was fully identical to session 1.

Trial exclusion

The trials in which participants failed the load and/or the deadline were excluded from subsequent analyses, since in these trials we could not ensure that deliberation was minimized during the initial stage. Participants failed to answer before the deadline on 4.6% of conflict initial trials (224 out of 4832) and 3.6% of no-conflict initial trials (175 out of 4832) of both test sessions combined. In addition, participants failed the load task on 14.9% of conflict initial trials (719 out of 4832) and 11.8% of no-conflict initial trials (572 out of 4832) of both test sessions. Overall, by rejecting the missed deadline and missed load trials, we kept 80.5% of conflict initial trials (3889 out of 4832) and 84.5% of no-conflict initial trials (4085 out of 4832) in session 1 and session 2 combined.

Conflict detection index

As mentioned before, conflict detection is typically calculated by subtracting the baseline confidence (i.e., the confidence at the correct no-conflict trials), from the confidence at the conflict trials (De Neys et al., 2013; Frey et al., 2018; Mevel et al., 2015; Pennycook et al., 2015). The higher the difference between the two, the more conflict is thought to be experienced by the participant. However, when reasoners deliberate on a problem, the initial doubt that they might have felt in relation to it can be dissolved (e.g., Bago & De Neys, 2020; De Neys et al., 2013). In this case, their reported confidence will not be a pure measure of the conflict that they initially experienced. To tackle this issue, previous one-response studies discarded correct conflict trials when calculating conflict detection, as in these trials the heuristic response had been overcome (i.e., the conflict associated with it had been resolved, e.g., De Neys & Glumicic, 2008; Pennycook et al., 2015; Šrol & De Neys, 2021). For the same reason, studies that use the two-response paradigm focus on the confidence of the initial responses for the

calculation of conflict detection. At this stage deliberation is experimentally minimized. Consequently, conflict detection at this stage gives a purer measure of intuitively experienced conflict, which should more directly reflect the strength of the posited intuitions (Bago & De Neys, 2017).

In addition, by using the confidence at the initial, intuitive trials, one can analyse both incorrect and correct conflict trials, since even correct trials will not be contaminated by deliberation. Note that participants still had to memorize the cognitive load pattern while providing their initial response confidence, which further ensured that their confidence was not affected by post-decision reflection.

Following the above studies and our preregistration, in the present paper we therefore focused on initial conflict detection. Response confidence was recorded both for the initial and the final responses, but we were a priori interested in the initial stage. Likewise, we only used confidence and not reaction times for the calculation of conflict detection, as the latter has been shown to be a less reliable indicator of detection ability (Frey et al., 2018; Šrol & De Neys, 2020), especially in a two-response setting (Bago & De Neys, 2017).

Finally, note that the rare trials in which no-conflict problems were solved incorrectly were discarded for the conflict detection analysis, since it is hard to interpret these unequivocally (see De Neys & Glumicic, 2008; Pennycook et al., 2015).

Composite Measure

For simplicity and to maximize power, our analyses focused on the composite level across the four individual reasoning tasks. To calculate this composite performance, for each participant, we calculated the proportion of correct initial and final responses for the conflict and no-conflict problems in each of the reasoning tasks and in each session. Then we averaged across all reasoning tasks (separately for each session, each response stage and conflict and no-conflict trials). For completeness, the individual task data is also included in our figures. Overall, the composite trends were reflected in the individual tasks.

Results

Statistical Analysis

The data were processed and analysed using the R software (R Core Team, 2020) and the following packages (in alphabetical order): dplyr (Wickham et al., 2021), ez (Lawrence, 2016), ggplot2 (Wickham, 2016), ggpubr (Kassambara, 2020), Rmisc (Hope, 2013), rstatix (Kassambara, 2021), and tidyr (Wickham, 2021).

Accuracy

To see if there was an effect of the response stage (initial; final) and the session (session 1; session 2) on the accuracy of conflict problems, a two-way within-subjects ANOVA was conducted. As Figure 1 shows, the accuracy at the conflict problems was significantly higher in the final than the initial response stage, $F(1, 150) = 11.07$, $p < .01$, $\eta^2g = 0.003$, which suggests that accuracy improved after deliberation. In addition, the accuracy at the conflict problems was significantly higher in session 2 compared to session 1, $F(1, 150) = 22.65$, $p < .001$, $\eta^2g = 0.01$, indicating that participants slightly improved when given a second chance to solve the problems. This improvement was independent of the response stage, as indicated by the lack of interaction between response stage and session, $F(1, 150) = 2.42$, $p = .12$; $\eta^2g < 0.001$.

As Figure 1 shows, these composite level trends were also observed for each individual task separately, with the exception of the conjunction fallacy problems in which final responses tended to be slightly less accurate than initial responses (see Dujmović et al., 2021, for a similar observation).

As expected, the average accuracy at the no-conflict problems remained at ceiling both for initial (M = 90.3, SD = 6.6 in session 1; M = 89.9, SD = 6.7 in session 2) and final responses (M = 92.9, SD = 7.0 in session 1; M = 92.3, SD = 7.6 in session 2), showing that participants paid attention throughout the study and refrained from guessing.

To summarize, although deliberation led to a slight improvement in performance, participants remained typically biased when solving classic conflict tasks. Overall, these results are in line with previous two-response studies (e.g., Bago & De Neys 2017, 2019a; Thompson et al., 2011).

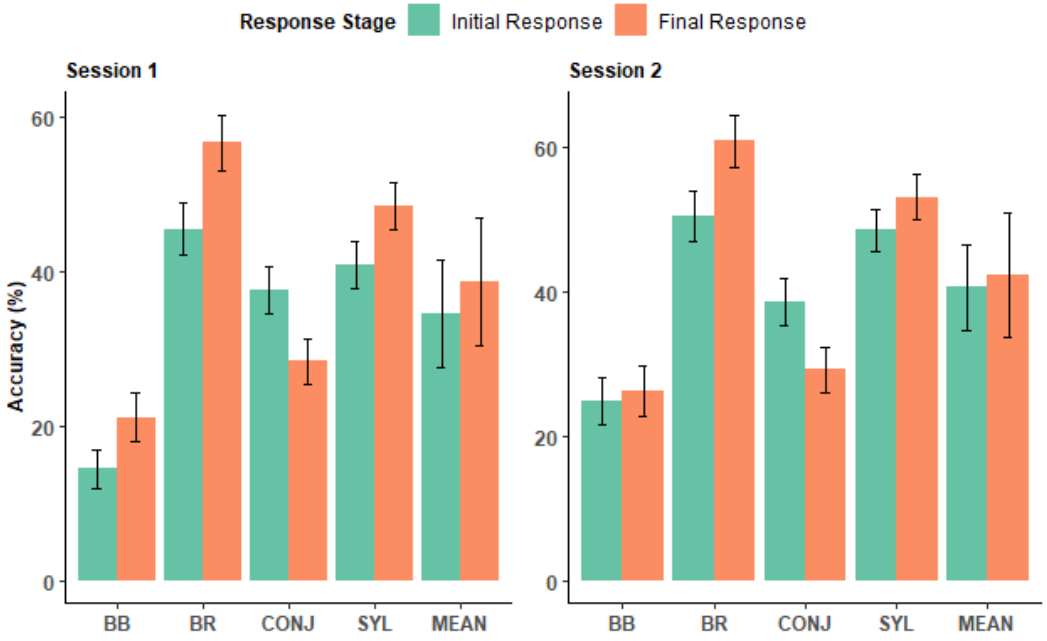

Figure 1. Proportion (%) of correct responses on the conflict problems, separately for each response stage, each session, each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

Direction of change

We also conducted a direction of change analysis on the conflict problems to explore whether and how participants changed their responses after deliberation (Bago & De Neys, 2017, 2019a). More specifically, we looked into how their accuracy changed (or did not change) from the initial to the final stage in every trial. At each response stage participants could either have an accuracy of "1" (i.e., correct response) or an accuracy of "0" (i.e., incorrect response). Since participants always provided two responses in a trial, we end up with four possible response patterns: "00" (incorrect initial and incorrect final response), "01" (incorrect initial and correct final response), "10" (correct initial and incorrect final response) and "11" (correct initial and correct final response). The results were consistent with previous findings (Bago & De Neys, 2017, 2019a). As Figure 2 shows, at the composite level, the majority of the conflict trials had a "00" pattern both in session 1 (54.8%) and in session 2 (52.2%), which confirms that reasoners are easily lured by the heuristic response when solving classic heuristics-and-biases tasks. We also note that, in the conflict trials, the proportion of "11" responses (28.0% in session 1; 35.1% in

session 2) was higher than that of the "01" responses (10.6% in session 1; 7.2% in session 2). As in previous two-response studies (e.g., Bago & De Neys, 2017, 2019a; Newman et al., 2017), this indicates that correct responses are, for the most part, already generated intuitively and not after deliberation. Finally, the least prevalent response pattern was "10" (session 1: 6.6%; session 2: 5.4%). As Figure 2 shows, these patterns were also observed on each of the individual tasks.

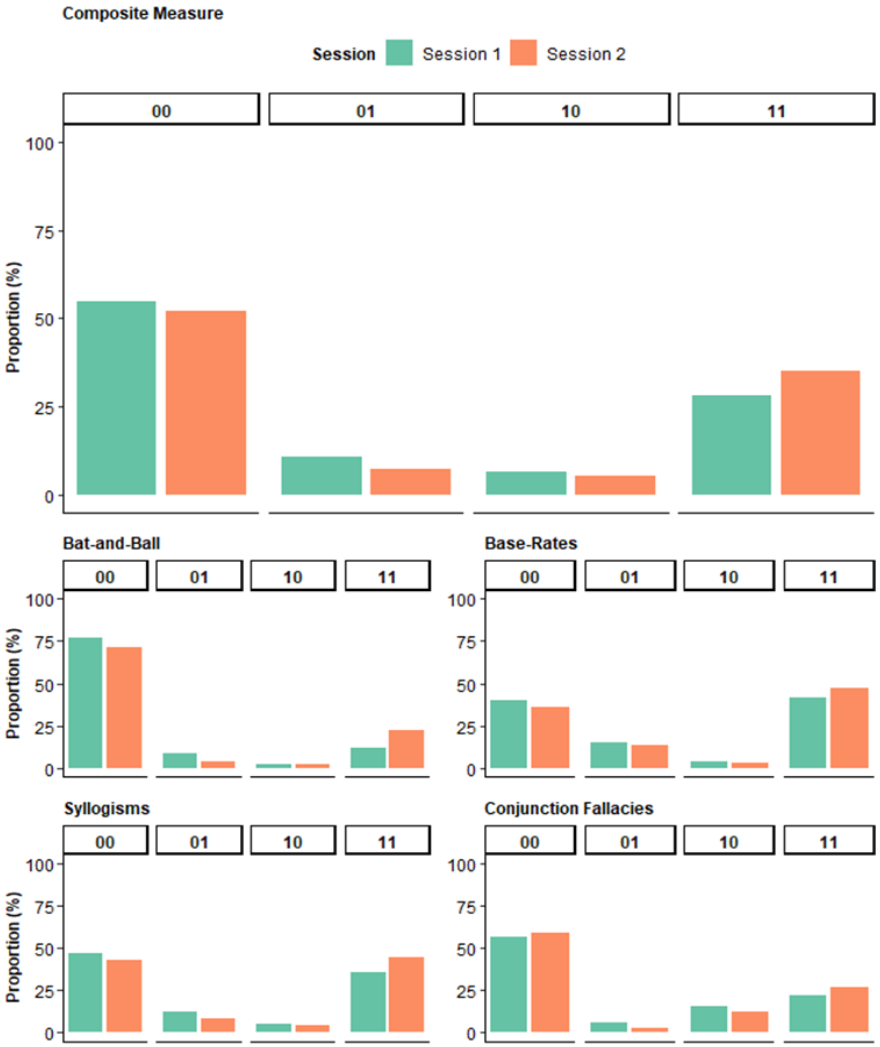

Figure 2. Proportion of each direction of change (i.e., "00" trials, "01" trials, "10" trials and "11" trials) for the conflict trials according to each session, each reasoning task, and the composite measure across the four reasoning tasks. "00" = incorrect initial and final response; "01" = incorrect initial and correct final response; "10" = correct initial and incorrect final response; "11" = correct final and correct initial response.

Accuracy Correlations

Before moving on to the core stability analyses we also examined whether the average accuracy of each individual at session 1 was correlated with the accuracy of that individual at session 2. A Pearson's product-moment correlation test revealed a high, positive accuracy correlation both for initial conflict trials, $r = 0.77$, $t(149) = 14.65$, $p < .001$, and for final conflict trials, $r = 0.84$, $t(149) = 18.73$, $p < .001$. The same pattern was observed for the individual tasks (see Supplementary Material Section A). Hence,

this indicates that those individuals who scored best the first time around, remained scoring well at the-retest. In this sense, the heuristics-and-biases tasks had a high test-re-test reliability.

Stability Index

Next, we investigated the stability of responses from session 1 to session 2. Stability is an inherently different measure of participants' responding than accuracy. To illustrate, consider an example of an exam with yes/no responses consisting of 20 items. The expected accuracy of an unprepared student is 50%. Now imagine that this student, still unprepared, had retaken the exam in the second term and always selected the opposite response compared to the first term. Their accuracy would still be 50%, but their stability would be 0%.

We separately calculated the stability of initial and final responses. Note that with respect to final responses, there are four possible patterns of (in)stability from session 1 to session 2: : "s00" (incorrect final response at both sessions), "s01" (incorrect final response at session 1 and correct final response at session 2), "s10" (correct final response at session 1 and incorrect final response at session 2), and "s11" (correct final response at both sessions). If the final response pattern of an individual item was "s00" or "s11", this item was categorized as "stable", whereas if the pattern was "s01" or "s10", the item was characterized as "unstable". The same stability classification was made for initial responses. These patterns should not be confused with the aforementioned direction of change patterns, hence the added "s", which stands for "stability". While the direction of change deals with the accuracy change from the initial to the final response of a trial, the direction of (in)stability deals with the accuracy change of a response (initial or final) from session 1 to session 2.

After all individual items were categorized as either stable or unstable, the average stability was calculated for each participant. As Figure 3A shows, we observed a very high stability both at the composite level and for each individual task, for the initial and final responses (initial response composite: M = 78.7%, SD = 17.2%; final response composite: M = 83.1%, SD = 14.6%). For completeness, note that we also observed the same pattern at the no-conflict trials (initial response composite: M = 90.2%, SD = 11.5%; final response composite: M = 93.3%, SD = 10.5%). This indicates that overall people's performance is highly stable after two weeks and reasoners rarely change their answers.

Direction of Change Stability. After establishing the stability of initial and final responses, we took a step further and examined the stability of the direction of change patterns from session 1 to session 2. More precisely, if a participant's trial had the same direction of change both in session 1 and session 2, this trial was coded as having a stable direction of change, and vice versa. We found that the stability of the direction of change category was high, both for conflict (M = 70.7%, SD = 19.7%) and no-conflict (M = 86.5%, SD = 15.1%) problems, which confirms that participants' response patterns were very consistent in time. More specifically, this finding indicates that, for the vast majority of the trials, the way people changed (or did not change) their initial responses after deliberation in session 1, was typically the way they changed them when re-tested two weeks later. As Figure 3B shows, the same trends were observed for the individual tasks.

However, at the same time it is clear that neither the responses nor the direction of change categories remained 100% stable from session 1 to 2, and we can still notice some response variability, especially so on the conflict problems. Our main aim was to see if conflict detection could explain this variability.

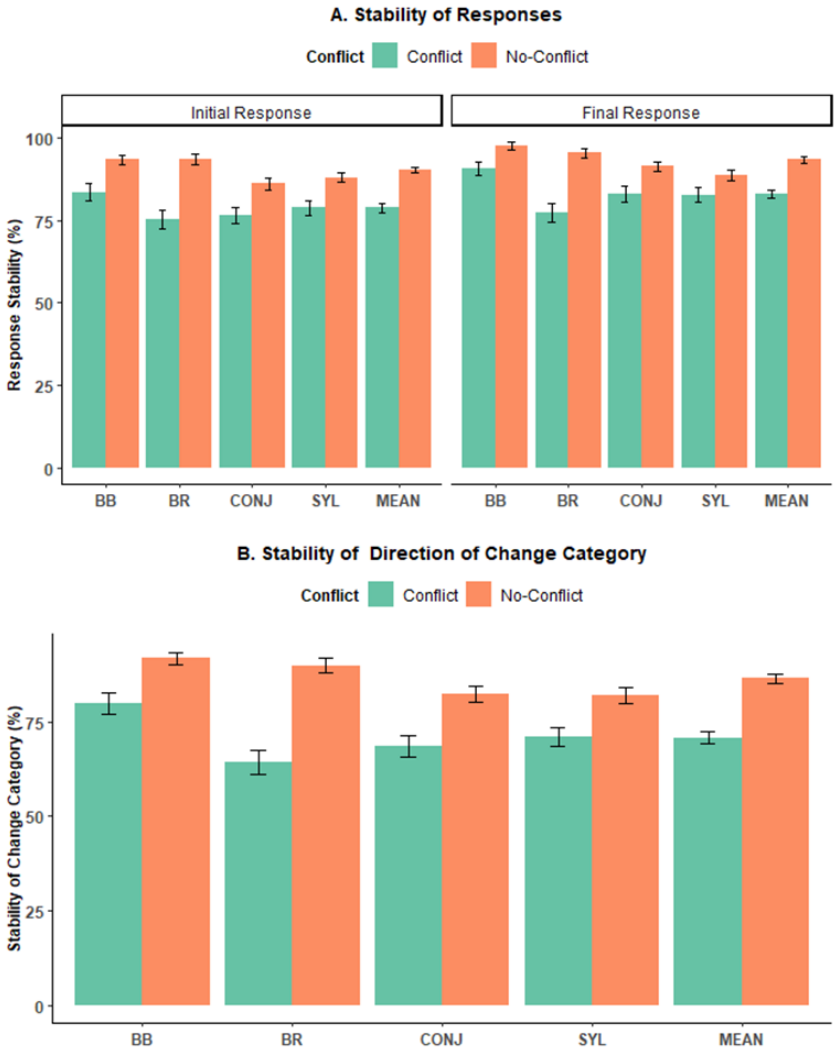

Figure 3. Panel A shows the proportion of responses that remained stable from session 1 to session 2, separately for conflict and no-conflict problems, for each response stage, each reasoning task and for the composite mean across the four tasks. Panel B shows the proportion of trials that had a stable direction of change category (i.e., "00" trials, "01" trials, "10" trials and "11" trials) from session 1 to session 2, separately for each response stage, each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

Conflict Detection

As a reminder, the conflict detection was calculated from the confidence ratings at the initial responses in the following manner: Confidence $_{conflict}$ – Confidence $_{no\text{-}conflict\_correct}$. For comparison with previous studies, we first wanted to check whether we observed an overall lower confidence on conflict versus no-conflict trials, pointing to a group-level conflict detection effect. This was indeed the case across tasks, responses and sessions (see Supplementary Material Section B). In addition, we also wanted to verify whether conflict detection was more pronounced on trials in which reasoners changed their response after deliberation ("01" and "10" trials), compared to trials in which reasoners did not change their response after deliberation ("00" and "11" trials, e.g., see Bago & De Neys, 2017, 2020;

Thompson et al., 2011)[3]. As Figure 4 shows, this pattern was consistently observed across tasks, responses and sessions. As in previous work, these results show that the higher the conflict experienced during an initial response, the more likely for this response to change in the final stage. Hence, both with respect to response accuracy and conflict detection, our findings are in line with previous two-response studies.

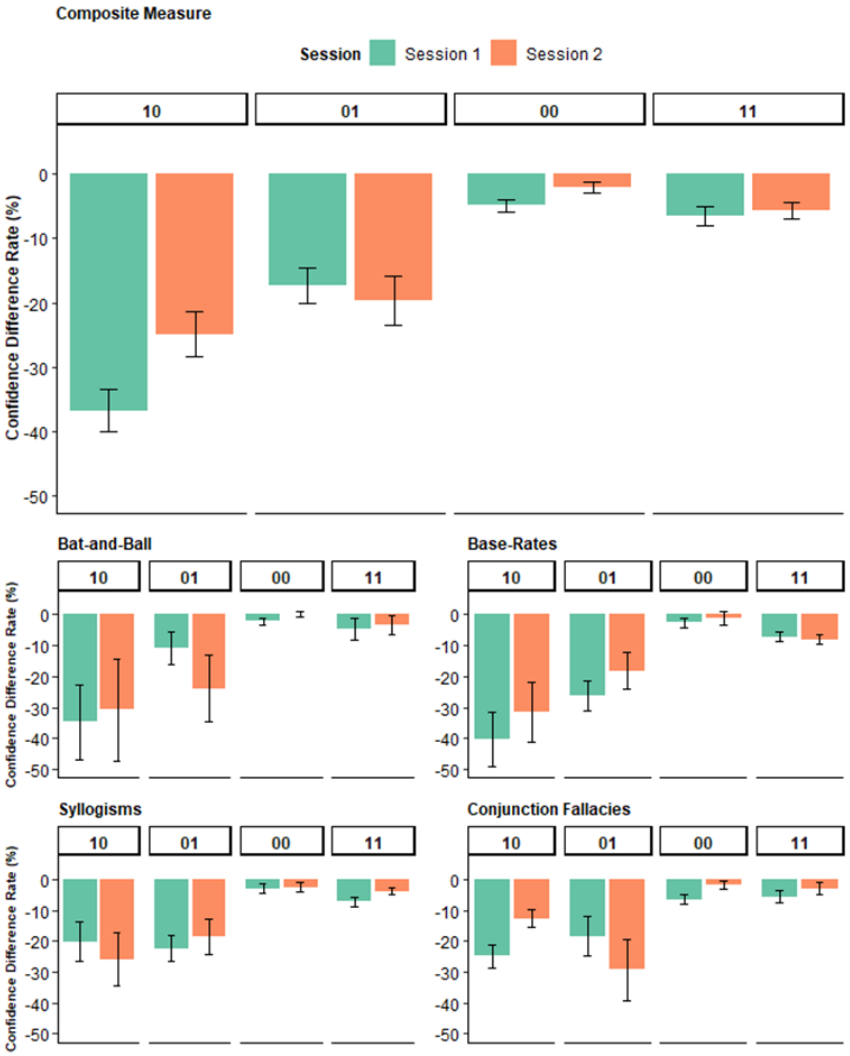

Figure 4. The mean confidence difference rate (%) according to the direction of change category (i.e., "01" trials and "10" trials represent the "change" categories, while "00" and "11" trials represent the "no change" categories), separately for each session, each reasoning task, and the composite measure across the four reasoning tasks. Negative values point to an overall successful conflict sensitivity. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

As one reviewer suggested, for comparison with previous one-response studies (e.g., De Neys & Glumicic, 2008; Pennycook et al., 2015; Šrol & De Neys, 2021), we re-ran this analysis by discarding

[3] Note that we used only the dominant no-conflict "11" category for this contrast, as responses in the other no-conflict direction of change categories cannot be interpreted unequivocally.

the correct conflict trials when calculating conflict detection (see Supplementary Material Figure S2 for the conflict detection means and Figure S3 for the conflict detection means for each direction of change category). Overall, the patterns and conclusions were consistent.

Predictive Conflict Detection

We now turn to the test of our main research question, in which we examine whether (initial) conflict detection at session 1 can predict response stability two weeks later, both at the intuitive and the deliberate level. In order to calculate conflict detection for every item of each participant, we first categorized the items of each participant as either "stable" or "unstable". More specifically, if a participant's accuracy at a given Item 1 was the same in session 1 and session 2, Item 1 would be classified as "stable" and vice versa. Once we classified all items as either "stable" or "unstable", we calculated, for each participant, the average conflict detection at all their stable items combined, and at all their unstable items combined. This way, each participant had two conflict detection indices: one for their stable and one for their unstable items. Inevitably, there were some participants whose items were all stable or all unstable throughout the study. Since these participants only had one conflict detection index (either for their stable or for their unstable items), they were examined separately. In the analyses below we were mainly interested in the composite measure and not the differences between the reasoning tasks. For completeness, we also report the data for each individual task. However, these individual task level analyses often have low sample sizes so they should be interpreted with some caution.

Note that as suggested by one reviewer, we also ran this analysis using the absolute confidence values at the initial conflict problem responses, also known as the feeling of rightness (Thompson et al., 2011), instead of the conflict detection indices (see Supplementary Material Figure S5 for the mean confidence values). As Supplementary Material Figure S6 shows, this type of analysis yielded the same pattern of results (see Supplementary Material Table S3 for the significance tests). In addition, we ran the same analysis after discarding the correct conflict trials when calculating conflict detection. As Figure S4 in the Supplementary Material shows, this analysis revealed the same results (see Supplementary Material Table S2 for the significance tests).

Initial Detection and Final Stability. By calculating the grand mean of conflict detection, we found that the initial detection was overall higher for the items that had unstable final responses (M = −9.9, SD = 11.6), compared to the initial detection of the items that had stable final responses (M = –7.3, SD = 9.6). This trend agrees with our hypothesis and, as Figure 5A shows, it is observed in all individual reasoning tasks. To test the statistical significance of these results we compared participants' composite conflict detection index at their stable and at their unstable items. Evidently, we only included the subjects that had both stable and unstable items (N = 114). Any participants with solely stable items were discarded from this analysis (there were no participants with only unstable items). A paired-samples t-test revealed a significant difference in the conflict detection indices between stable (M = –5.6, SD = 11.3) and unstable (M = –12.0, SD = 22.1) items; $t(113) = 3.05$, $p < .01$. As expected, the unstable items had a higher conflict detection compared to the stable ones. It is worth noting that participants with only stable items (N = 37), had a very low average conflict detection (M = –3.8, SD = 6.9).

Initial Detection and Initial Stability. Next, we performed the same analysis as above, but now we focused on how initial conflict detection impacted the initial, intuitive responses at session 2. Consistent

with the above results, we found that the grand mean of the composite conflict detection index was overall higher for the items with unstable initial responses (M = –19.1, SD = 20.7), compared to the conflict detection of the items with stable initial responses (M= –6.0, SD = 8.8). As Figure 5B shows, this trend was observed on all individual reasoning tasks. To test the statistical significance of these results we compared participants' composite conflict detection index at their stable and at their unstable items. Evidently, we only included the subjects that had both stable and unstable items (N = 122). Any participants with solely stable items were discarded from this analysis (there were no participants with only unstable items). A paired samples t-test revealed a significant difference in the conflict detection scores between stable (M = –3.5, SD = 7.8) and unstable (M = –18.5, SD = 26.1) items, $t(121) = 6.19$, $p < .001$. It is worth noting that participants that had only stable items (N = 29), had a low average conflict detection (M = –3.2, SD = 7.2).

Our main a priori conflict detection measure concerned the detection at the initial, intuitive response level. For exploratory purposes, we repeated the analysis, this time using the conflict detection at the final responses as a predictor of (initial and final) response stability. Supplementary Material Figure S7 shows the results. Although the trends tended to be slightly weaker, overall the same pattern was observed, in that unstable trials showed a more pronounced conflict detection than stable trials.

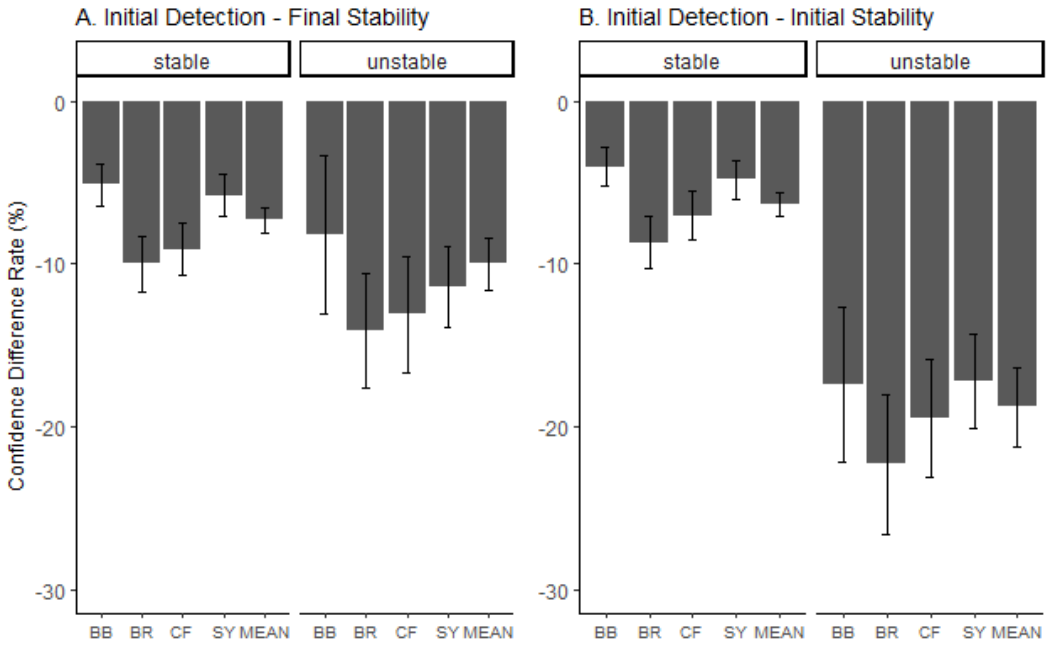

Figure 5. The grand means of the initial conflict detection index (i.e., Confidence $_{conflict}$ – Confidence $_{no\text{-}conflict\_correct}$) according to stability (stable; unstable). Panel A shows the average initial conflict detection according to the stability of the final responses and Panel B shows the average initial conflict detection according to the stability of the initial responses, separately for each reasoning task and for the composite mean across the four tasks. Negative values point to an overall successful conflict sensitivity. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CF = Conjunction Fallacies; SY = Syllogisms; MEAN = the composite mean across the four tasks.

Discussion

In the present paper we focused on the temporal stability of reasoning performance and examined a potential determinant for answer change. Participants solved the same tasks twice in two

test sessions, two weeks apart. We used the two-response paradigm to test the stability of both initial (intuitive) and final (deliberate) responses. We hypothesized that participants who showed higher conflict detection in their initial, intuitive responses at session 1, would be less stable in their responses between session 1 and session 2. Conflict detection was operationalized as the confidence difference for initial responses on classic conflict problems versus control no-conflict problems.

Results point to two main conclusions. First, people's responses to classic "bias" tasks are highly stable. In general, participants rarely changed their intuitive and deliberate answers two weeks after they were first tested. This result is in line with the findings by Białek and Pennycook (2018) and Stango and Zinman (2020) who—in one of the rare direct tests of the stability of heuristics-and-biases tasks—also observed that individual biases remained highly stable over time. From a psychometric perspective, the high stability of the performance on heuristics-and-biases tasks is obviously excellent news. This is particularly important as the performance on these tasks is frequently used in the literature as a predictor of a wide range of variables (e.g., Baron et al., 2015; Białek & Sawicki, 2018; Shenhav et al., 2012; Toplak et al., 2017; West et al., 2008). If people's task performance would not be stable, this would undermine its use as a predictor. In this sense, the findings validate the popular use of these tasks by showing that they exhibit an adequate test-retest reliability.

Second, despite the high stability, there was still some variability in initial and final responses after the first test. By directly comparing the conflict detection for items that had a stable accuracy to those that had an unstable accuracy, we found that the initial conflict detection was significantly higher in the unstable items. In other words, the higher the initial conflict detection participants experienced on an item, the more likely they were to change their responses to this item two weeks later. This finding indicates that the variability of responses over time is not entirely random, but can be predicted (Stango & Zinman, 2020).

At the methodological level, we believe that the current findings further underline the potential of the two-response paradigm (Thompson et al., 2011), which has become increasingly popular in the past years (e.g., Bago & De Neys, 2017, 2019a, 2021; Burič & Konradova, 2021; Burič & Srol, 2020; Dujmovic et al., 2021; Vega et al., 2021). As we have mentioned, previous work showed that conflict detection can predict answer change on an intra-trial level. Conflict detection during the initial stage is much more pronounced in the cases that participants change their initial answers during the final response stage (Bago & De Neys 2017, 2019a; Thompson & Johnson, 2014). With the present study, we show that conflict detection at the initial stage does not only predict answer change in the short, intra-trial term, but also in the longer term, between separate test sessions. The generalization of the conflict detection and answer change coupling over a longer time window points to an interesting new application of the paradigm.

At the theoretical level, conflict detection (or a lowered feeling of rightness in the conceptualization of Thompson et al., 2011) is often conceived as a triggering mechanism that allows a reasoner to switch from System 1 intuiting to System 2 deliberation (e.g., De Neys, 2012; Pennycook et al., 2015; Thompson et al., 2011). One consequence of engaging in deliberation is that people might revise their intuitively generated answer (Thompson et al., 2011). With respect to the stability of final responses, this suggests that conflict experienced at time 1 will make it more likely that the reasoner engages in deliberation at time 1, but also at time 2, two weeks later. Because deliberation increases the probability of answer change, it will be more likely that reasoners give a different final response at time 1 and time 2.

But interestingly, our findings not only concerned the final but also at the initial responses. By definition, in the initial response stage deliberation is minimized and, hence, answer change cannot be

driven by differential deliberation per se. So why does conflict detection predict initial answer stability? Our hypothesis was inspired by recent advances in dual process theorizing in which the intuitive reasoning performance is determined by the strength interplay of competing intuitions (e.g., Bago & De Neys, 2020; De Neys & Pennycook, 2019; Pennycook et al., 2015). As we noted, these models postulate that the "logical" response that has traditionally been considered to be cued by System 2, can also be cued by System 1. Hence, it is assumed that when reasoners are faced with a traditional heuristics-and-biases task, System 1 will not only give rise to the traditionally postulated "heuristic" intuition, but also to a "logical" intuition (which is assumed to be based on automatically activated learned mathematical and probabilistic rules, e.g., De Neys, 2012). Whichever intuition is strongest will be selected as initial response. The more similar the strength of the competing intuitions, the more conflict will be experienced. If one intuition clearly dominates over the other, the dominant intuition will be generated with little or no experienced conflict. We reasoned that any accidental noise at different test sessions will be more likely to affect (revert) the strength ordering of competing intuitions that showed little differentiation to start with. Going back to our introductory analogy, the clearer your preference for one dessert over another, the more likely that you will make the same choice repeatedly. Hence, a highly dominant intuition (indexed by low conflict detection) will be more likely to remain dominant at re-test than a less dominant intuition (indexed by high conflict detection). Consequently, conflict detection will also predict answer stability of the intuitive response.

Obviously, this theoretical account remains speculative. The strength of competing intuitions is a hypothetical construct and was not directly measured. We also acknowledge that this construct can be defined in various ways (e.g., processing "fluency" or "speed"). At present, the specific processes underlying the relationship between logical and heuristic intuitions have not been specified, and we do recognise the need for their precise implementation.

It is worth noting that the current findings are also relevant for the discussion on Individual Differences in conflict detection. Previous studies have shown that, although most people might detect the conflict in their answers, not everyone does (e.g., Frey et al., 2018; Pennycook et al., 2015; Šrol & De Neys, 2021). The high response stability in our study and its relation to a low conflict detection, suggests that there are some participants who always remain biased and unaware of their errors. In other words, some reasoners consistently provide incorrect answers (i.e., they do not change their erroneous responses at time 2) and they have low or no conflict detection at time 1.

One may also note that the observed high stability of participants' responses, both on the intra-trial level and between the separate test sessions, suggests that most participants respond on an intuitive basis even when they are given the time to deliberate. However, we would like to highlight that this does not imply that deliberation is never used or needed when it comes to sound reasoning. Although response change was rare in our study, there were still cases in which people engaged in deliberation to correct their intuitive answers (i.e., "01" cases). In addition, recent studies have suggested that deliberation might be helpful to provide explicit justifications for an intuitive insight (see Bago & De Neys, 2019a; De Neys & Pennycook, 2019).

It is clear that the approach we introduced here can be further developed and fine-tuned. For example, for practical reasons (e.g., attrition) the present study focused on a two week time window. This presents a dramatic departure from the millisecond intra-trial time-scale that two-response studies typically focus on to study answer change. But, obviously, one could further expand the timeline and test the predictability of answer stability at time points that are months or even years apart. Likewise, the present study has focused on heuristics-and-biases tasks only. The two-response paradigm has been used to explore answer change in different domains (e.g., moral reasoning, Bago & De Neys, 2019b;

Vega et al., 2021; or prosocial reasoning in economic settings, Bago et al., 2021). In theory, the present approach can be adopted to test the predictability of long-term answer change in all these fields.

To conclude, the present study showed that people's responses to heuristics-and-biases tasks are highly stable. The rare cases in which answers are nevertheless changed seem to be driven by the detection of conflict between competing intuitions. We believe that the results point to the potential of the approach and hope that it can inspire new applications in the reasoning and decision-making fields.

Acknowledgements

This research was supported by a research grant (DIAGNOR, ANR16-CE28-0010-01) from the Agence Nationale de la Recherche, France. Aikaterini Voudouri was supported by an IdEx doctoral fellowship from the University Paris Cité.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Open Data Statement

Data for this study are publicly available at OSF: https://doi.org/10.17605/OSF.IO/QZM2T

ORCID

Aikaterini Voudouri https://orcid.org/0000-0001-7415-7631

References

Bago, B., Bonnefon, J. F., & De Neys, W. (2021). Intuition rather than deliberation determines selfish and prosocial choices. *Journal of Experimental Psychology: General*, *150*(6), 1081–1094. https://doi.org/10.1037/xge0000968

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

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

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

Bago, B., & De Neys, W. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, *4*(3), 265–284. https://doi.org/10.1016/j.jarmac.2014.09.003

Białek, M., Muda, R., Stewart, K., Niszczota, P., & Pieńkosz, D. (2020). Thinking in a foreign language distorts allocation of cognitive effort: Evidence from reasoning. *Cognition, 205*, 104420. https://doi.org/10.1016/j.cognition.2020.104420

Białek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior research methods, 50*(5), 1953-1959. https://doi.org/10.3758/s13428-017-0963-x

Białek, M., & Sawicki, P. (2018). Cognitive reflection effects on time discounting. *Journal of Individual Differences. 39*(2), 99–106 https://doi.org/10.1027/1614-0001/a000254

Boissin, E., Caparos, S., Raoelison, M., & De Neys, W. (2021). From bias to sound intuiting: Boosting correct intuitive reasoning. *Cognition*, *211*, 104645. https://doi.org/10.1016/j.cognition.2021.104645

Burič, R., & Konrádová, Ľ. (2021). Mindware Instantiation as a Predictor of Logical Intuitions in the Cognitive Reflection Test. *Studia Psychologica*, *63*(2), 114–128. https://doi.org/10.31577/sp.2021.02.822

Burič, R., & Šrol, J. (2020). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. *Journal of Cognitive Psychology*, *32*(4), 460–477. https://doi.org/10.1080/20445911.2020.1766472

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. https://doi.org/10.1177/1745691611429354

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, *20*(2), 169–187. https://doi.org/10.1080/13546783.2013.854725

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

De Neys, W. (2021). On dual-and single-process models of thinking. *Perspectives on psychological science, 16*(6), 1412-1427. https://doi.org/10.1177/1745691620964172

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299. https://doi.org/10.1016/j.cognition.2007.06.002

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/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. https://doi.org/10.3758/s13423-013-0384-5

Dujmović, M., Valerjev, P., & Bajšanski, I. (2021). The role of representativeness in reasoning and metacognitive processes: An in-depth analysis of the Linda problem. *Thinking & Reasoning*, *27*(2), 161–186. https://doi.org/10.1080/13546783.2020.1746692

Epstein, S. (1994). Integration of the Cognitive and the Psychodynamic Unconscious. *American Psychologist*, *49*(8), 709–724. https://doi.org/10.1037/0003-066X.49.8.709

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

Evans, J. S. B., & Over, D. E. (2013). *Rationality and reasoning.* Psychology Press. https://doi.org/10.4324/9780203027677

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

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. https://doi.org/10.3758/s13421-016-0618-7

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology*, *71*(5), 1188–1208. https://doi.org/10.1080/17470218.2017.1313283

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—In search of a phenomenon. *Thinking & Reasoning*, *21*(4), 383–396. https://doi.org/10.1080/13546783.2014.980755

Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 28–43. https://doi.org/10.1037/a0021098

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

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

Kassambara, A. (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

Kassambara, A. (2021). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. R package version 0.7.0. https://CRAN.R-project.org/package=rstatix

Lawrence, M. A. (2016). ez: Easy Analysis and Visualization of Factorial Experiments. R package version 4.4-0. https://CRAN.R-project.org/package=ez

Mata, A. (2020). Conflict detection and social perception: bringing meta-reasoning and social cognition together. *Thinking & Reasoning*, *26*(1), 140-149. https://doi.org/10.1080/13546783.2019.1611664

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

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, *27*(2), 227–237. https://doi.org/10.1080/20445911.2014.986487

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

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

Newman, I., Gibb, M., & Thompson, V. (2017). Rule-Based Reasoning Is Fast and Belief-Based Reasoning Can Be Slow: Challenging Current Explanations of Belief-Bias and Base-Rate Neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1154–1170. https://doi.org/10.1037/xlm0000372

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition*, *42*(1), 1–10. https://doi.org/10.3758/s13421-013-0340-7

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, *124*(1), 101–106. https://doi.org/10.1016/j.cognition.2012.04.004

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive psychology, 80*, 34-72. https://doi.org/10.1016/j.cogpsych.2015.05.001

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL 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.

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

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423–428. https://doi.org/10.1037/a0025391

Sloman, S. A. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological bulletin, 119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Šrol, J., & De Neys, W. (2021). Predicting individual differences in conflict detection and bias susceptibility during reasoning. *Thinking & Reasoning*, *27*(1), 38–68. https://doi.org/10.1080/13546783.2019.1708793

Stango, V., & Zinman, J. (2020). Behavioral Biases are Temporally Stable. *Unpublished working paper.* https://doi.org/10.3386/w27860

Stupple, E. J. N., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking & Reasoning*, *19*(1), 54–77. https://doi.org/10.1080/13546783.2012.735622

Stupple, E. J., Gale, M., & Richmond, C. R. (2013). Working memory, cognitive miserliness and logic as predictors of performance on the cognitive reflection test. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the cognitive science society* (pp. 1396–1401). Austin, TX: Cognitive Science Society.

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

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

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-World Correlates of Performance on Heuristics and Biases Tasks in a Community Sample: Heuristics and Biases Tasks and Outcomes. *Journal of Behavioral Decision Making*, *30*(2), 541–554. https://doi.org/10.1002/bdm.1973

Trippas, D., & Handley, S. (2017). The parallel processing model of belief bias: Review and extensions. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 28–46). Oxon, UK: Routledge.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Vega, S., Mata, A., Ferreira, M. B., & Vaz, A. R. (2021). Metacognition in moral decisions: Judgment extremity and feeling of rightness in moral intuitions. *Thinking & Reasoning*, *27*(1), 124–141. https://doi.org/10.1080/13546783.2020.1741448

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, *100*(4), 930–941. https://doi.org/10.1037/a0012842

Wickham, H. (2016*). ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H. (2021). tidyr: Tidy Messy Data. R package version 1.1.3. https://CRAN.R-project.org/package=tidyr

Wickham, H., François, R., Henry, L., & Müller, K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.6. https://CRAN.R-project.org/package=dplyr

Supplementary Material

A. Accuracy Correlations

Table S1.

Pearson's product-moment correlation tests between the average accuracy of each individual at the conflict problems of session 1, and the accuracy of that individual at the conflict problems of session 2, separately for each reasoning task.

Response stage	Task	r	df	t
Initial response	BB	0.69	143	11.37*
	BR	0.67	140	10.76*
	SYL	0.65	145	10.44*
	CONJ	0.62	146	9.45*
Final response	BB	0.84	143	18.41*
	BR	0.65	140	10.14*
	SYL	0.71	145	12.09*
	CONJ	0.68	146	11.29*

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; CONJ = Conjunction Fallacies.

* $p < .001$.

B. Conflict Detection

As it can be seen in Figure S1 (note that negative values point to an overall successful conflict sensitivity) participants detected the conflict of their answers both at the initial and the final response stages, both at session 1 (initial: M = –7.0, SD = 8.6; final : M = –6.7, SD = 8.4) and session 2 (initial: M = –3.7, SD = 6.5; final: M = –3.6, SD = 6.2). The overall individual conflict detection at session 1 was significantly correlated with that of session 2 at the initial responses ($r = 0.32$, $t(149) = 4.08$, $p < .001$), but not at the final responses ($r = 0.27$, $t(149) = 3.41$, $p < .001$).

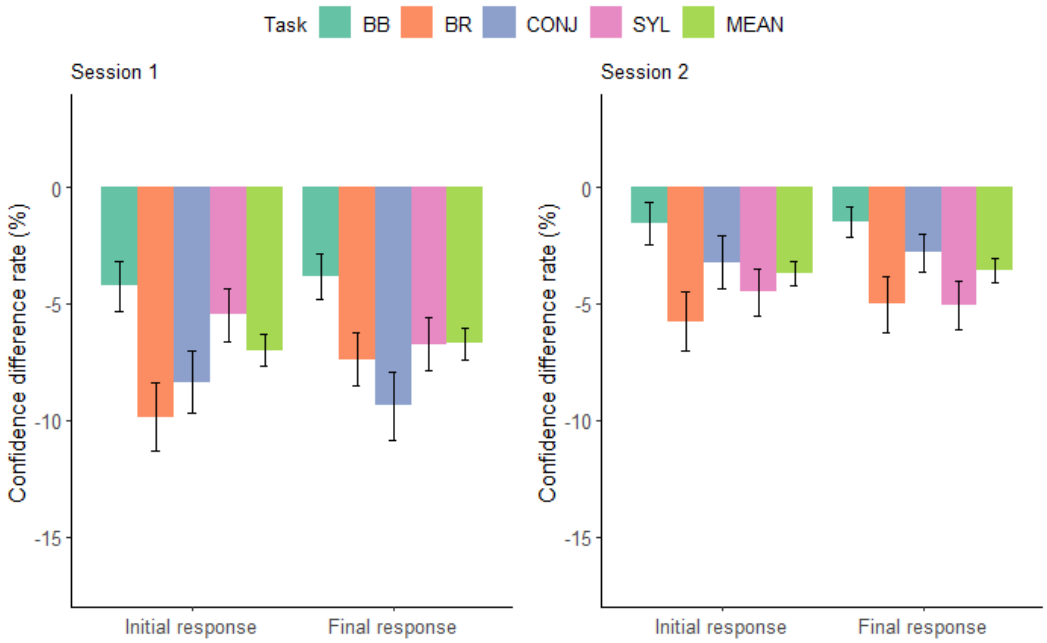

Figure S1. Confidence difference rates (%) between the conflict trials and the correct no-conflict trials (i.e., Confidence $_{conflict}$ – Confidence $_{no\text{-}conflict\_correct}$), separately for each session, each response stage, each reasoning task and for the composite mean across the four tasks. Negative values point to an overall successful conflict sensitivity. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

C. (Predictive) Conflict Detection on Incorrect Conflict trials

For completeness, in this section we re-ran the conflict detection and predictive conflict detection analyses by discarding the correct conflict trials when calculating conflict detection (i.e., conflict detection = Confidence $_{conflict\_incorrect}$ – Confidence $_{no\text{-}conflict\_correct}$). Due to the exclusion of incorrect conflict trials, we could only focus on the "00" and "01" directions.

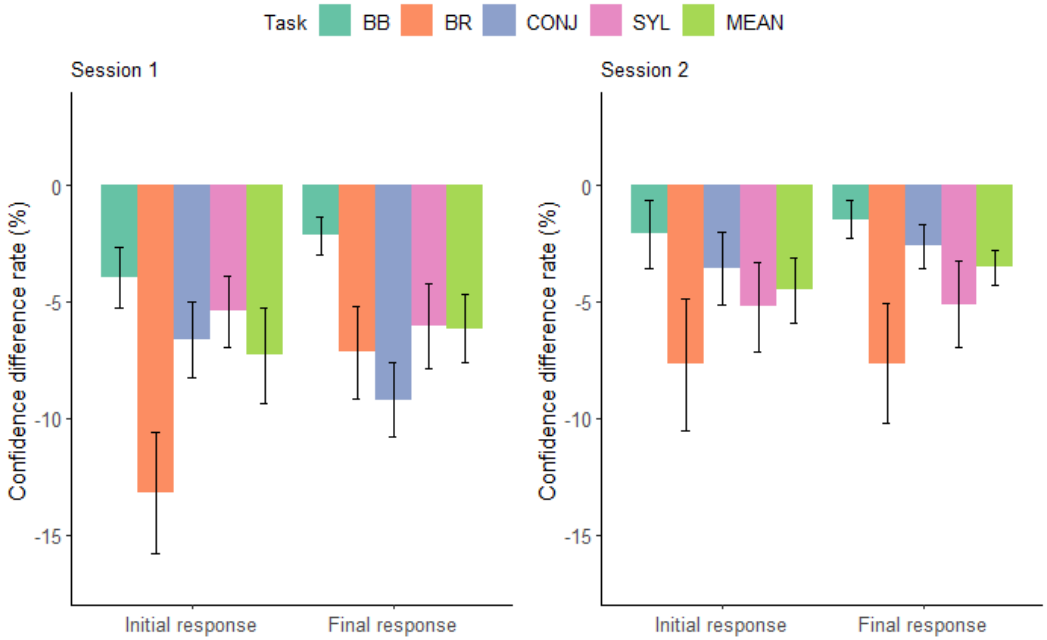

Figure S2. Confidence difference rates (%) between the incorrect conflict trials and the correct no-conflict trials, separately for each session, each response stage, each reasoning task and for the composite mean across the four tasks. Negative values point to an overall successful conflict sensitivity. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

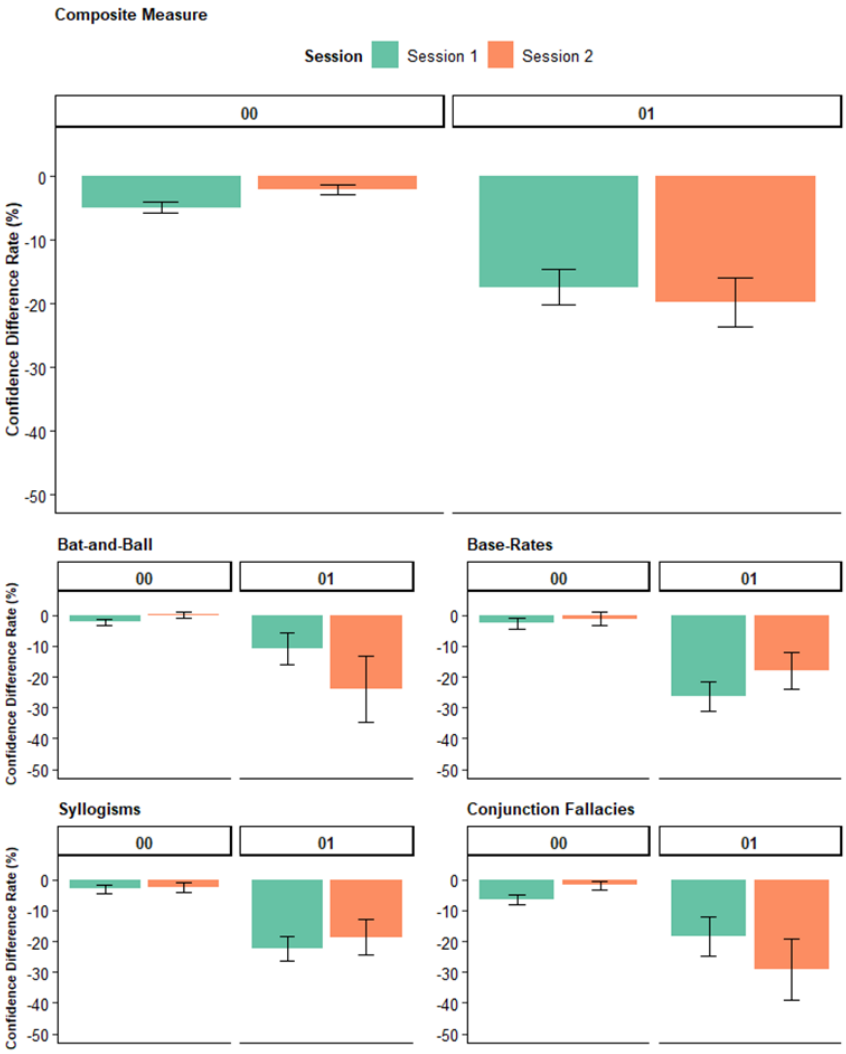

Figure S3. The mean confidence difference rate (%) according to the direction of change category (i.e., "01" trials represent the "change" category, "00" trials represent the "no change" category), separately for each session, each reasoning task and the composite measure across the four reasoning tasks. Negative values point to an overall successful conflict sensitivity. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

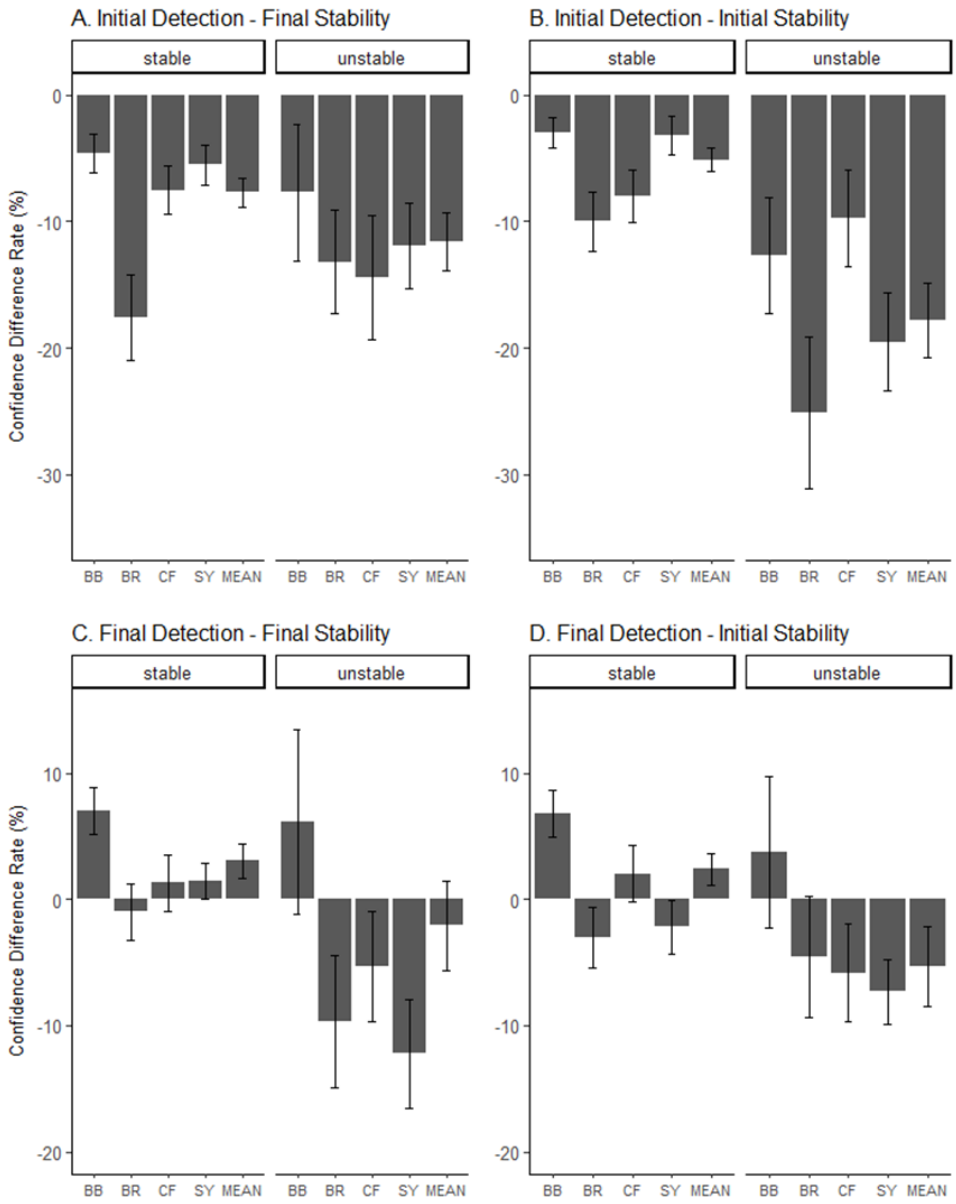

Figure S4. The (initial and final) conflict detection (i.e., Confidence $_{conflict\_incorrect}$ – Confidence $_{no\text{-}conflict\_correct}$) grand means according to stability (stable; unstable). Negative values point to an overall successful conflict sensitivity. Panel A shows the average initial conflict detection according to the stability of the final responses, Panel B shows the average initial conflict detection according to the initial responses' stability, Panel C shows the average final conflict detection according to the final responses' stability, and Panel D shows the average final conflict detection according to the initial responses' stability, separately for each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CF = Conjunction Fallacies; SY = Syllogisms; MEAN = the composite mean across the four tasks.

Table S2.

Paired-samples t-tests between the mean conflict detection of the stable items and the mean conflict detection of the unstable items of each individual.

		Mean (SD) stable	Mean (SD) unstable	t	df
Initial detection	Final stability	–5.6 (13.8)	–11.6 (22.4)	2.36*	87
	Initial stability	–3.8 (10.3)	–14.6 (25.9)	3.95***	94
Final detection	Final stability	4.8 (14.8)	–6.5 (21.1)	3.88***	76
	Initial stability	3.7 (17.6)	–3.5 (19.5)	2.71**	89

* $p < .05$.

** $p < .01$.

*** $p < .001$.

D. (Predictive) Confidence Values

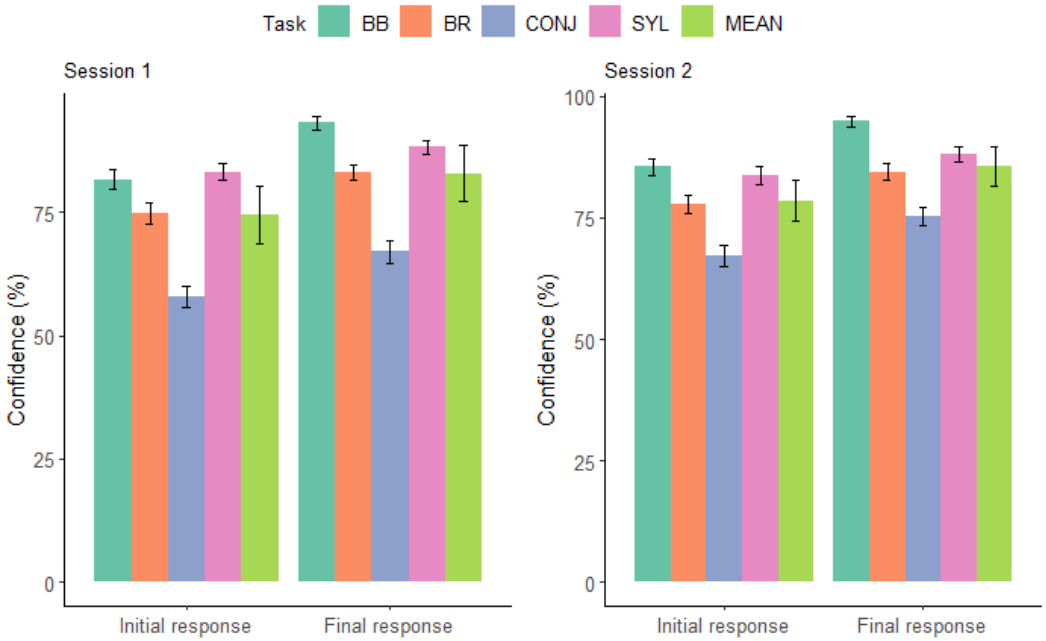

Figure S5. Confidence rates (%) at the conflict trials, separately for each session, each response stage, each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CONJ = Conjunction Fallacies; SYL = Syllogisms; MEAN = the composite mean across the four tasks.

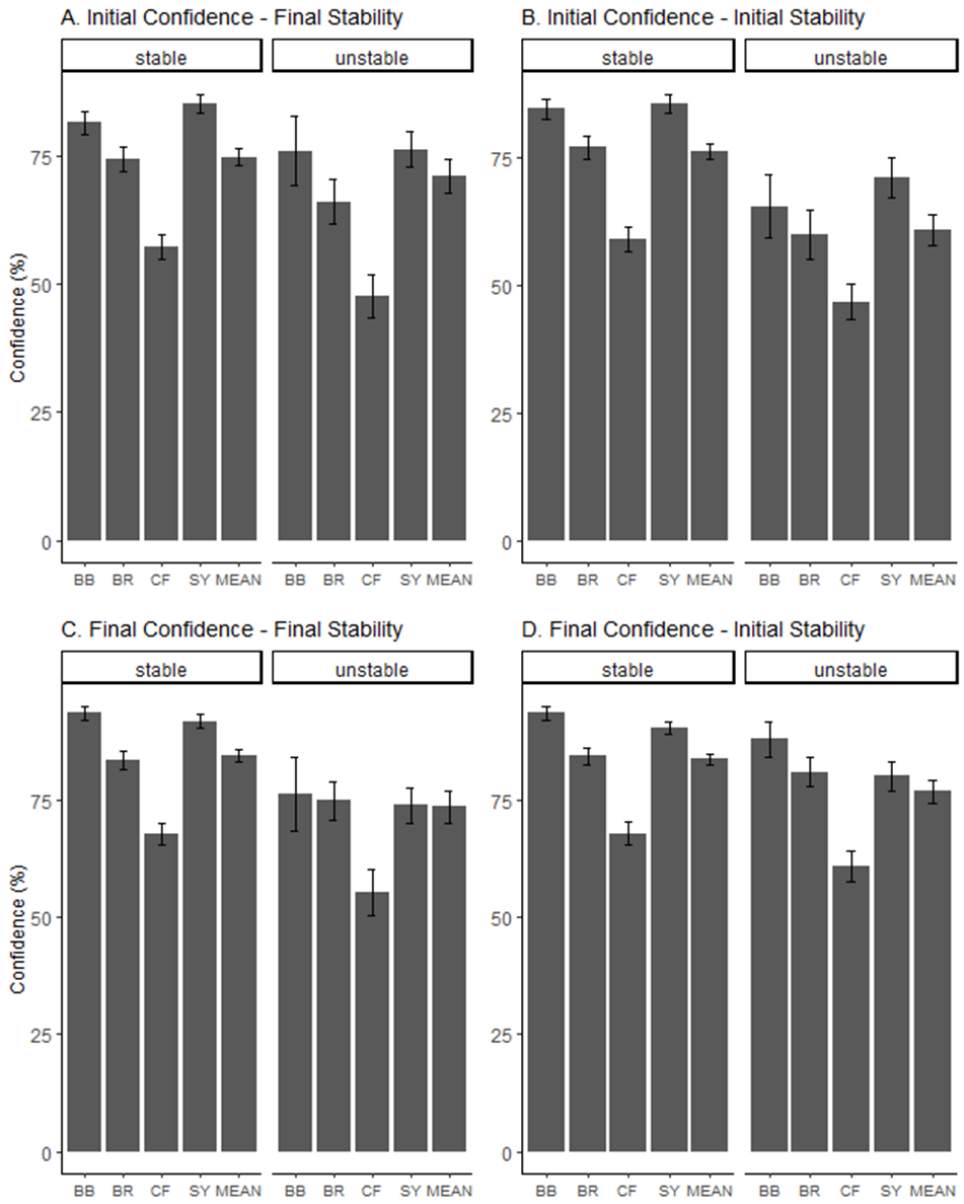

Figure S6. The (initial and final) confidence grand means according to stability (stable; unstable). Panel A shows the average initial confidence according to the stability of the final responses, Panel B shows the average initial confidence according to the initial responses' stability, Panel C shows the average final confidence according to the final responses' stability, and Panel D shows the average final confidence according to the initial responses' stability, separately for each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CF = Conjunction Fallacies; SY = Syllogisms; MEAN = the composite mean across the four tasks.

Table S3.

Paired-samples t-tests between the mean confidence of the stable items and the mean confidence of the unstable items of each individual.

		Mean (SD) stable	Mean (SD) unstable	t	df
Initial confidence	Final stability	76.6 (24.6)	65.1 (29.4)	4.47*	115
	Initial stability	78.4 (22.3)	60.2 (31.1)	6.80*	123
Final confidence	Final stability	87.1 (18.9)	69.1 (30.1)	6.88*	115
	Initial stability	87.4 (18.1)	76.6 (24.8)	5.98*	123

* $p < .001$.

E. Predictive Conflict Detection of Final Responses

Final Detection and Final Stability. By calculating the grand mean of conflict detection at the final responses, we found that there was a conflict detection effect for the items that had unstable final responses (M = –7.5, SD = 20.4), but a lack of conflict detection effect for the items with stable final responses (M = 2.3 , SD = 12.8), as indicated by the positive confidence difference between conflict and no-conflict trials. As Figure S7A shows, this trend is observed in most individual reasoning tasks. To test the statistical significance of these results we compared participants' composite (final) conflict detection index at their stable and at their unstable items. Evidently, we only included the subjects that had both stable and unstable items (N = 114). Any participants with solely stable items were discarded from this analysis (there were no participants with only unstable items). A paired-samples t-test revealed a significant difference in the final conflict detection indices between stable (M = 3.3, SD = 14.1) and unstable (M = –9.1, SD = 24.5) items; $t(113) = 4.89$, $p < .001$. As expected, the unstable items had a higher conflict detection compared to the stable ones. It is worth noting that participants with only stable items (N = 37), did not show a conflict detection effect (M = 3.6, SD = 6.4).

Final Detection and Initial Stability. By calculating the grand mean of conflict detection at the final response, we found that there was a conflict detection effect for the items that had unstable initial responses (M = –2.3 , SD = 15.8), but no conflict detection effect for the items that had stable initial responses (M = 1.4, SD = 11.9). As Figure S7B shows, this trend is observed in most individual reasoning tasks. To test the statistical significance of these results we compared participants' composite conflict detection index at their stable and at their unstable items. Again, we only included the subjects that had both stable and unstable items (N = 122). Any participants with solely stable items were discarded from this analysis (there were no participants with only unstable items). A paired-samples t-test revealed a significant difference in the conflict detection indices between stable (M = 3.5, SD = 10.9) and unstable (M = –3.2, SD = 18.6) items; $t(121) = 4.09$, $p < .001$. As expected, the unstable items had a higher conflict detection compared to the stable ones. Like in the above analysis, participants with only stable items (N = 29), did not show a conflict detection effect (M = 2.3, SD = 6.9).

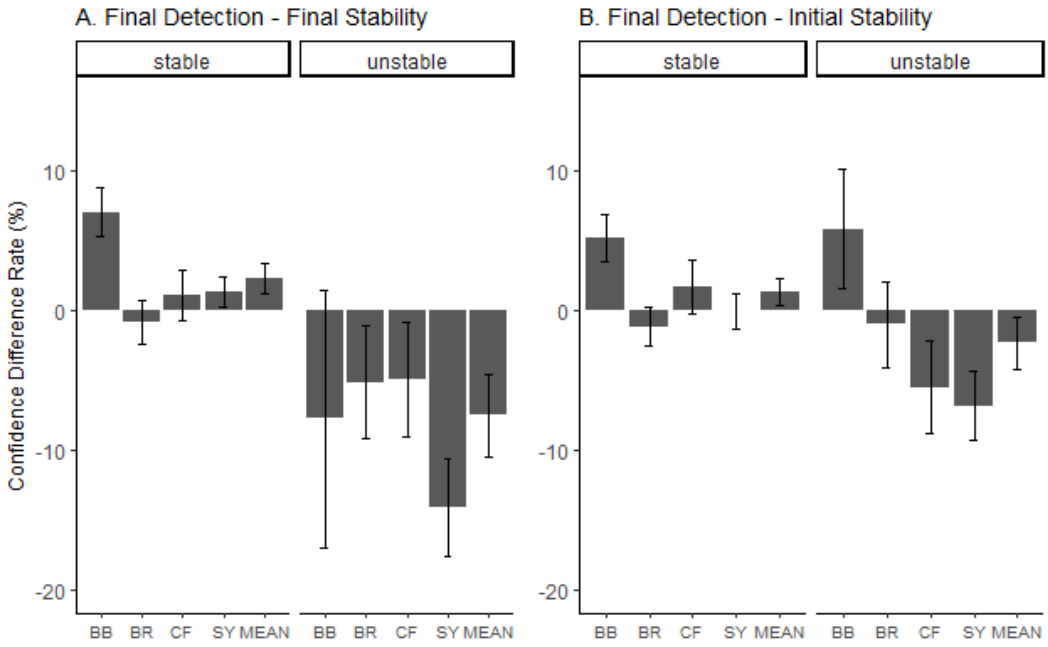

Figure S7. The grand means of the final conflict detection index (i.e., Confidence $_{conflict}$ – Confidence $_{no\text{-}conflict\_correct}$) according to stability (stable; unstable). Negative values point to an overall successful conflict sensitivity. Panel A shows the average final conflict detection according to the stability of the final responses and Panel B shows the average final conflict detection according to the stability of the initial responses, separately for each reasoning task and for the composite mean across the four tasks. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; CF = Conjunction Fallacies; SY = Syllogisms; MEAN = the composite mean across the four tasks.