Judgment and Decision Making (2023), **Vol. 18**:e33 1–31
doi:10.1017/jdm.2023.32

EMPIRICAL ARTICLE

Reasoning and cognitive control, fast and slow

Aikaterini Voudouri[1], Bence Bago[2,3], Grégoire Borst[1] and Wim De Neys[1]

[1]Universite Paris Cite, LaPsyDE, CNRS, Paris, France; [2]Artificial and Natural Intelligence Toulouse Institute, University of Toulouse, Toulouse, France and [3]Institute for Advanced Study in Toulouse, University of Toulouse, Toulouse, France

Corresponding author: Aikaterini Voudouri; Email: aikaterini.voudouri@gmail.com

Received: 15 July 2023; **Revised:** 1 August 2023; **Accepted:** 1 August 2023

Keywords: dual-process theory; cognitive control; Stroop task; Flanker task; reasoning; two-response paradigm

Abstract

Influential 'fast-and-slow' dual process models suggest that sound reasoning requires the correction of fast, intuitive thought processes by slower, controlled deliberation. Recent findings with high-level reasoning tasks started to question this characterization. Here we tested the generalizability of these findings to low-level cognitive control tasks. More specifically, we examined whether people who responded accurately to the classic Stroop and Flanker tasks could also do so when their deliberate control was minimized. A two-response paradigm, in which people were required to give an initial 'fast' response under time–pressure and cognitive load, allowed us to identify the presumed intuitive answer that preceded the final 'slow' response given after deliberation. Across our studies, we consistently find that correct final Stroop and Flanker responses are often non-corrective in nature. Good performance in cognitive control tasks seems to be driven by accurate 'fast' intuitive processing, rather than by 'slow' controlled correction of these intuitions. We also explore the association between Stroop and reasoning performance and discuss implications for the dual process view of human cognition.

1. Introduction

Sometimes a solution to a problem pops into mind instantly and effortlessly whereas at other times arriving at a decision can take time and effort. This distinction between what is often referred to as a more intuitive and deliberate mode of cognitive processing—or the nowadays more popular 'System 1' and 'System 2' labels—lies at the heart of the influential 'fast-and-slow' dual process view that has been prominent in research on human reasoning in the last decades (Evans, 2008; Kahneman, 2011).

Although intuitive thinking is useful when it comes to fast decision-making, it often also relies on mental shortcuts, or heuristics, which can lead to cognitive biases (Kahneman, 2011). This bias susceptibility of System 1 is often demonstrated in the literature with the use of heuristics-and-biases tasks, like the following example:

> A psychologist wrote thumbnail descriptions of a sample of 1000 participants consisting of 5 women and 995 men. The description below was drawn randomly from the 1000 available descriptions.
>
> Sam is a 25 years old writer who lives in Toronto. Sam likes to shop and spends a lot of money on clothes.

What is most likely?

a. Sam is a woman.

b. Sam is a man.

Intuitively, many people will be tempted to conclude that Sam is a woman based on stereotypical beliefs cued by the description. However, given that there are far more males than females in the sample (i.e., 995 out of 1000), the statistical base rates favor the conclusion that a randomly drawn individual will most likely be a man. Hence, logically speaking, taking the base rates into account should push the scale to the 'man' side. Unfortunately, educated reasoners are typically tricked by their intuition and often fail to solve the problem correctly (e.g., De Neys and Glumicic, 2008).

The dual process framework presents a simple and elegant explanation for this bias phenomenon (Evans, 2008; Kahneman, 2011). Dual process theorists have traditionally highlighted that taking logical principles into account typically requires demanding System 2 deliberation (e.g., Evans, 2002, 2008; Evans and Over, 1996; Kahneman, 2011; Stanovich and West, 2000). Because human reasoners have a strong tendency to minimize difficult computations, they will often refrain from engaging or completing the slow deliberate processing when mere intuitive processing has already cued a response (Evans and Stanovich, 2013; Kahneman, 2011). Consequently, most reasoners will simply stick to the intuitive response that quickly came to mind and fail to consider the logical implications. It will only be the few reasoners who have sufficient resources and motivation to complete the deliberate computations and override the initially generated intuitive response, who will manage to reason correctly and give the logical answer (Stanovich and West, 2000). Hence, sound reasoning is, in essence, believed to be corrective in nature.

However, studies in the last decade suggest we may need to reconsider this traditional view of the two systems (De Neys and Pennycook, 2019). These studies typically present heuristics-and-biases tasks using a two-response paradigm (Thompson et al., 2011). More specifically, participants are asked to provide two consecutive responses on each task trial. The first response is given under time–pressure and a cognitive load (e.g., a parallel task taxing cognitive resources), while in the final response stage participants have no restrictions and are allowed to deliberate (Bago and De Neys, 2017). Since System 2 is believed to be slow and burden our cognitive resources, the constraints that are imposed during the initial response minimize its involvement. This way, the paradigm allows for a direct comparison of more intuitive and deliberate responses. The key finding of these studies is that in many of the (infrequent) trials where participants provide a correct, final response, they had already provided a correct response during the initial stage (e.g., Bago and De Neys, 2017, 2019a; Newman et al., 2017; Raoelison and De Neys, 2019). Hence, System 2 does not always need to revise the intuitively generated responses, as the latter might already be correct.

Relatedly, a similar line of research using the two-response paradigm has shown that when people provide biased intuitive responses, they are often sensitive to the fact that they are erring (De Neys, 2017; Pennycook et al., 2015). In other words, participants seem not completely oblivious to the fact that their answers conflict with some (logical) elements of the problem. This has been found by comparing conflict/incongruent and no-conflict/congruent versions of the same heuristics-and-biases tasks. In congruent versions, both the heuristic and logical information in the problem cue the same answer. For instance, the congruent version of the example given above would simply switch the base rates around (e.g., 'A psychologist wrote thumbnail descriptions of a sample of 1000 participants consisting of 995 women and 5 men'). Everything else stays the same. Hence, in the congruent case, both the description and the base-rates cue the same response (i.e., 'Sam is a woman'). If processing logical principles such as base-rate information requires deliberation, then reasoners' initial, intuitive responses to the incongruent and the congruent versions should not differ. However, when solving incongruent trials, participants typically report lower confidence in their initial responses in comparison to congruent trials. This response doubt has been referred to as conflict detection in the reasoning field and suggests that participants are intuitively processing the conflicting information in the incongruent

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

problem (e.g., Bago and De Neys, 2017; Burič and Šrol, 2020; Mata, 2020; Pennycook et al., 2014; Thompson and Johnson, 2014; but also Mata et al., 2014; Mata and Ferreira, 2018).

The above findings have led researchers to propose a revised dual process model–sometimes referred to as a 'Dual Process model 2.0'—which posits that System 1 can generate 2 types of intuitions, a classic 'heuristic' intuition, and an alleged 'logical' intuition (e.g., Bago and De Neys, 2017, 2019a; De Neys and Pennycook, 2019; Handley et al., 2011; Newman et al., 2017; Pennycook et al., 2015; see De Neys, 2017, for review). The latter is believed to be based on an automated knowledge of mathematical and probabilistic rules (De Neys, 2012; Evans, 2019; Stanovich, 2018).

Interestingly, similar patterns have also been observed in other higher-order reasoning tasks on moral (Bago and De Neys, 2019b; Vega et al., 2021) and prosocial (Bago et al., 2021; Kessler et al., 2017) reasoning. The main result across these studies is that responses that are assumed to require deliberation by the traditional dual-process model (e.g., taking the consequences of a moral action into account or maximizing pay-offs for oneself or others), are often generated intuitively. It then seems that there is a need to upgrade our view of the fast and intuitive System 1. Responses that are traditionally believed to necessitate controlled deliberation, often seem to fall within the realm of more intuitive processing (De Neys, 2022; De Neys and Pennycook, 2019).

The key aim of the present article is to explore the generalizability of these findings to classic cognitive control tasks, like the Stroop task (Stroop, 1935) and the Flanker task (Eriksen and Eriksen, 1974). These are tasks that have been used to directly tap into lower-level cognitive control processes, rather than higher-order functioning, such as reasoning. Cognitive control, according to a common definition, is a group of top-down processes that help us carry out cognitive tasks when automatic responding is not sufficient (Botvinick et al., 2001; Diamond, 2013). Similar to heuristics-and-biases tasks, classic cognitive control tasks usually contain 2 competing pieces of information: task-relevant and task-irrelevant information. In the incongruent versions, the task-irrelevant information cues an automatic, incorrect response, which conflicts with the response cued by the task-relevant information. Conversely, in the congruent version, both the task-relevant and task-irrelevant information cue the same response.

For example, one of the most popular and frequently used tasks is the Stroop (Stroop, 1935). In the Stroop, task participants are presented with words that denote a color and are written in a colored ink (e.g., the word 'red' written in blue ink). Sometimes the ink color and the word are congruent (e.g., the word 'red' written in red ink), but other times, as in the first example, they are incongruent. Participants are asked to respond to the ink color of each word. On average, participants have longer reaction times and higher error rates when solving the incongruent compared to the congruent stimuli. This is also known as the Stroop interference effect. The most common explanation for this effect is that, since reading is an automatic process for educated adults, reading the word will always come before identifying its ink color (Keele, 1972; Stirling, 1979; but also Kahneman and Chajczyk, 1983). Therefore, in the incongruent trials, participants need to take the time to inhibit their automatically generated (incorrect) answer (i.e., the read word), in order to arrive at the correct answer (i.e., the ink color in which the word is written). In other words, not giving in to the luring, automatic response is thought to require controlled, effortful processing (e.g., Botvinick et al., 2001). Put differently, cognitive control is assumed to have a corrective role: fast (incorrect) responses are generated automatically and are then corrected by slower controlled processes. This pattern is similar to the one that has been put forward by traditional dual process theories in the reasoning field: heuristic responses are generated automatically, and are later corrected by slow, deliberate processes (e.g., Evans and Stanovich, 2013). However, as it was mentioned before, the corrective role of deliberation in the reasoning field has been questioned, and evidence shows that correct responses are often generated automatically. Given the reasoning findings, our goal in the present article is to examine whether correct responding to cognitive control tasks is also possible when control is minimized.

It is worth mentioning that, in line with this research question, recent cognitive control findings have shown evidence for an automatically operating (cognitive) control (Desender et al., 2013; Jiang et al., 2015, 2018; Linzarini et al., 2017). These studies focus on a phenomenon observed in cognitive control

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

tasks, where participants tend to more often respond correctly to an incongruent trial if it is preceded by an incongruent trial (instead of a congruent one, e.g., Braem and Egner, 2018). The explanation for this phenomenon is that the cognitive control that is recruited during the first trial facilitates correct responding in the upcoming trial. Critically, studies have found that this effect persists even when the first trial is presented unconsciously (e.g., Desender et al., 2013; Jiang et al., 2015, 2018; Linzarini et al., 2017). This suggests that cognitive control on the subliminal trial can, in theory, be exerted automatically (without the participants' intention, e.g., Abrahamse et al., 2016; Algom and Chajut, 2019). These findings lend some credence to the idea that correct responding in cognitive control tasks might be observed in the absence of deliberate correction.

In Studies 1–3 of the present article, we directly tested this hypothesis and examined whether correct responding in cognitive control tasks is also possible when participants' deliberate control is constrained. For this purpose, we focused on the Stroop task (Studies 1 and 3) and the Flanker task (Study 2). In the Flanker task participants were presented with a central arrow surrounded by 2 arrows on each side. The surrounding arrows either pointed in the opposite direction (incongruent trials) or the same direction (congruent trials) as the central arrow (Stoffels and Van der Molen, 1988) and participants' task was to indicate the direction of the central arrow. We designed a two-response version of both the Stroop task and the Flanker task. We were specifically interested in testing whether, in the incongruent trials where participants managed to provide a correct final response, they had already arrived at a correct response in the initial stage or not.

A second objective of the present article (Study 3), was to explore in what way cognitive control and reasoning performance are related. There is existing evidence in the literature showing that classic cognitive control tasks can predict reasoning accuracy (Abreu-Mendoza et al., 2020; De Neys et al., 2011; Handley et al., 2004). Participants who score better on cognitive control tasks, such as the Stroop, tend to show less biased responding on reasoning tasks.

Despite the links between these 2 measures, the way in which they are related is unclear. If we assume that correct responding in both cognitive control and reasoning tasks results from the same generic mechanism, we can imagine (at least) 2 possible alternative routes. On the one hand, it could be that both reasoning and cognitive control tasks tap onto the same deliberate control processes. In other words, people who successfully control (and later correct) their automatically generated Stroop responses, would also be good at controlling (and correcting) their intuitive responses in reasoning tasks. Under this 'smart deliberator' view (Raoelison et al., 2020), people's performance in the Stroop task would predict their ability to deliberately correct responses in reasoning tasks. On the other hand, it might also be the case that both reasoning and cognitive control tasks tap into intuitive or automatic control processes. In other words, people who provide correct Stroop responses when their cognitive resources are restricted would also be able to intuitively provide correct responses to reasoning problems. When these people are allowed to deliberate, they will not need to correct their intuitive answers, as these will be already correct. Under this 'smart intuitor' view, people's 'intuitive' performance at the Stroop task would predict their ability to generate correct intuitive responses (rather than to deliberately correct their intuitions) in the Reasoning task. In Study 3, we presented participants with both a two-response Stroop task and a set of two-response reasoning tasks to explore this issue.

2. Study 1

In Study 1, we designed a two-response version of the Stroop task. On each trial, participants were asked to give a first answer as fast as possible under cognitive constraints (time–pressure and secondary memorization task load) and to then take the time to reflect and provide a final constraint-free response. The key question is whether correct responding to the critical incongruent Stroop trials is also possible when participants' deliberate control is constrained. In those cases that participants managed to provide a correct final response, do they initially typically err or is the initial response already correct?

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

2.1. Method

2.1.1. Preregistration and data availability

The study design and hypothesis were preregistered on the Open Science Framework (https://osf.io/9pz5j). No specific analyses were preregistered. All data and material are also available on the Open Science Framework (https://osf.io/gkhbm/).

2.1.2. 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. Participants were paid £2.60 for their participation (£5 hourly rate). Based on Aïte et al.'s (2016) Stroop study, we recruited 50 adult participants. The mean age of participants was 37.2 years (SD = 14.3) and 60% were female. Thirty-four percent of participants had a high-school degree as their highest education level, 50% had a bachelor's degree, 12% had a Master's degree, and 4% had not completed high school.

2.1.3. Materials

2.1.3.1. Stroop stimuli

Based on Aïte et al. (2016), 16 color-word stimuli were created by combining 4 different color names ('red', 'green', 'blue', and 'yellow') with 4 corresponding ink colors (RGB color codes 255;0;0, 0;255;0, 0;0;255 and 255;255;0). We used these stimuli to create 64 congruent and 64 incongruent Stroop experimental trials. Before the main experiment, participants were presented with a set of practice trials (Section 2.1.4.2). For the color practice, 4 circle stimuli were created each filled with either red, green, blue, or yellow ink (RGB color codes 255;0;0, 0;255;0, 0;0;255 and 255;255;0).

All stimuli were presented in the center of the screen on a gray background (RGB code 135; 135; 135) in randomized order. Participants were instructed to press the key 'd' if the word was presented in the color red, the key 'f' if it was presented in blue, the key 'j' if it was presented in green, and the key 'k' if it was presented in yellow (we chose these 4 response keys as they have the same position in the 3 most common keyboard layouts: QWERTY, QWERTZ, and AZERTY). The response times were measured from the stimulus onset until the button press.

Congruent trials allowed us to test for a guessing confound and are reported in this context. Our main results concern the critical incongruent trials unless otherwise stated.

2.1.3.2. Load task

In the two-response version of the Stroop task (Section 2.1.4.2), we used a secondary digit memorization task (Lavie, 2005; Lavie et al., 2004), as this type of task has been shown to burden cognitive control in Stroop-like tasks (i.e., it has been found that this task increases the Stroop interference effect, e.g., de Fockert et al., 2001; Lavie and De Fockert, 2005; Lavie et al., 2004; but also Gao et al., 2007). On each trial, participants saw a sequence of 6 (black) digits (i.e., the memory set). All digits were randomly selected from 1 to 9 without replacement on a given trial. The memory probe consisted of a single black digit, a question mark, and a message reminding participants of the keyboard response buttons. Participants were asked to indicate whether the probe had appeared in the memory set on that trial. They were instructed to press 'd' for probe-present and 'k' for probe-absent responses. For half of the trials, the correct answer was 'probe present'.

2.1.4. Procedure

2.1.4.1. One-response (deliberative-only) pre-test

In order to obtain a baseline Stroop performance, we conducted a pre-test where participants performed a traditional one-response color-word Stroop task, without a digit memorization load or a deadline. We recruited 25 participants (52% female; mean age = 35.4 years, SD = 17.3) online on Prolific Academic (www.prolific.ac). Only native English speakers from Canada, Australia, New Zealand, the United

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

States of America, or the United Kingdom were allowed to take part in the study. Participants were paid £1.00 for their participation. A total of 40% of the participants reported a high-school degree as their highest education level, while 56% reported a bachelor's degree and 4% a Master's degree.

The idea was to base the response deadline of the initial response stage in our two-response design on the average response time in the one-response pretest (e.g., Bago and De Neys, 2017, 2020). Thus, in the one-response pretest participants were presented with the same amount of trials and the same stimuli as in the main two-response study. The only difference from the main study was that participants were asked to provide a single response and they only received standard Stroop task instructions to respond 'as fast and as accurate as possible'. The average response time for the congruent trials was 755 ms (SD = 134 ms) and for the incongruent trials, it was 893 ms (SD = 197 ms).[1] Based on these values we decided to set the maximum response deadline for the initial response to 750 ms (i.e., approximately the mean of congruent trials which do not require controlled processing to answer correctly).

To verify that participants were indeed under time pressure during the initial stage, we compared the response times for the critical incongruent trials between the one-response pre-test and the initial responses in the main two-response study. For this comparison, we excluded all trials with missed load memorization or missed deadlines in the initial stage of the two-response study. The results showed that participants responded much faster in the initial response stage of the main study (incongruent trials: 580.8 ms, SD = 55.4 ms), compared to that of the one-response study (incongruent trials: 893.3 ms, SD = 196.5 ms; i.e., responses were on average more than 1.5 SDs faster than in the one-response study). A Welch Two Sample *t*-test indicated that this difference was significant ($t(26.47) = 7.76$, $p < .001$).

In addition, the one-response pre-test allowed us to check for a potential consistency confound in our main two-response study. More specifically, since the study requires two consecutive responses, participants might provide the same response in the initial and the final stage, merely driven by the desire to appear consistent (Thompson et al., 2011). In this case, the correction rate from the initial to the final response would be underestimated. Previous two-response work in other fields has argued against the presence of this confound (Bago and De Neys, 2017, 2019a, 2020; Thompson et al., 2011). Here we tested for it by contrasting the proportion of correct responses in the incongruent Stroop trials of the one-response pretest and those of the final stage of the main two-response study. A consistency confound would result in a clear discrepancy between these accuracies. However, our results showed that the percentage of correct responses in the critical incongruent trials of the one-response pretest (M = 93.2%, SD = 18.2%), was very similar to this of the incongruent final responses of the two-response study (M = 92.5%, SD = 18.6%). A Welch Two Sample *t*-test indicated that this difference was not significant ($t(52.32) = 0.14$, $p = 0.890$).

2.1.4.2. Two-response Stroop task

The experiment was run online on Gorilla Experiment Builder (gorilla.sc). Participants were informed that the study would take 30 minutes to complete and that it demanded their full attention. They were told that they would be presented with words and that they needed to respond to the color that each word was presented in using their keyboard (for literal instructions, see Supplementary Material Section A). Then they were given instructions about the correct response key mapping.

To familiarize themselves with the color-key pairs, participants first practiced only with the colors (without the words). They were presented with 32 color stimuli (red, blue, green, or yellow) and they were instructed to respond as fast and as accurately as possible. They were given feedback after each response and, in case of an incorrect response, they were shown a picture of a keyboard with the correct color-key pairs. Figure 1A illustrates the time course of this practice round.

[1] Before computing the average reaction times all trials with reaction times higher than 2 SDs above the general mean were removed from the analysis.

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

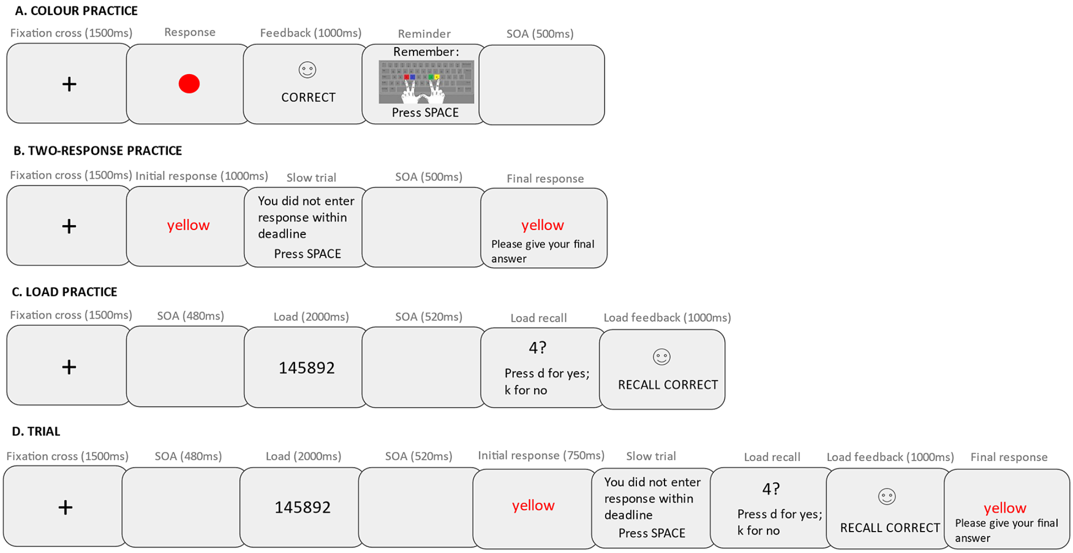

Figure 1. *Time course of the practice trials and experimental trial. (A) The time course of a color-only practice trial. (B) The time course of a deadline-only two-response practice trial. (C) The time course of a load-only practice trial. (D) The time course of an experimental trial.*

Then, participants were presented with a second practice round, which was identical to the first one, with the difference that now the stimuli were 12 congruent color-word pairs. Participants were told that they needed to respond to the color that each word was presented in.

After the second practice round, participants were introduced to the incongruent trials. They were informed that sometimes the ink color in which the word appears would not match with the word, and they were asked to always respond to the color of the word. This practice round was identical to the above 2, with the difference that now the stimuli were 8 incongruent color-word pairs.

At the end of this practice round, participants were introduced to the two-response paradigm. They were told that we were interested in their initial, intuitive response to the color of each word and wanted them to answer as fast as possible with the first response that popped up in mind. They were also informed that after the first response, they would have more time to reflect on the color of the word and provide their final answer. Participants were introduced to the deadline of the initial response and were shown an example of an initial trial. Then, they were presented with 12 two-response color-word trials. The time course of this practice round can be seen in Figure 1B.

Following the two-response practice round, participants were presented with the load task. They were told that they also had to memorize a set of 6 numbers while responding to the color-word pairs. Participants were informed that after the memory probe was shown, they would have to press 'd' if the probe was part of the memory set, or 'k' if the probe was not part of the memory set. At this point, they were presented with 5 load memorization practice trials. Figure 1C illustrates the time course of this practice round.

After the load practice, round participants were reminded that they had to memorize the set of numbers while responding to the color-word pairs. They were instructed to first focus on the memorization task, and then on the color-word task. They were then presented with 24 two-response practice trials (with load and deadline). Critically, the first 12 practice trials had a looser initial response deadline (1 second instead of 750 ms). This was done to familiarize participants with the two-response format. For the last 12 practice trials the actual 750 ms deadline was applied. The time course of this practice round was identical to that of the experimental trials and is illustrated in detail in Figure 1D.

After this practice session, participants started the experimental trials. The main task was composed of 128 trials which were grouped into 3 blocks. Participants were told that after each block they could take a short break. Before each new block started, they were shown a picture of a keyboard with the correct color-key pairs to remind them of the response key mapping. At the end of the experiment, participants completed standard demographic questions and were presented with a debriefing message.

2.1.5. Exclusion criteria

Following our preregistration, we discarded from all analyses participants who scored lower than 50% on both their initial congruent and initial incongruent trials. This was done to sidestep the possibility that results would be distorted because some participants could not meet the initial trial constraints without guessing. Based on this criterion, 6 out of the 50 participants were excluded. We were thus left with a sample of 44 participants (59% female) with a mean age of 36.6 years (SD = 14.1).

In addition, we excluded the trials in which participants failed the load and/or the deadline, since in these trials we could not ensure that deliberation was minimized during the initial stage. Participants failed to answer before the deadline on 36.2% of incongruent initial trials (1019 out of 2816) and 25.4% of congruent initial trials (716 out of 2816). In addition, participants failed the load task on 9.6% of incongruent initial trials (269 out of 2816) and 12.6% of congruent initial trials (355 out of 2816). Overall, we kept 58.1% of all trials (3273 out of 5632), by rejecting trials in which participants missed the deadline and failed the load task. On average, each participant contributed 74.4 trials (out of 128 trials, SD = 39.2). Clearly, the high amount of missed trials demonstrates that meeting the initial deadline and load constraints was challenging for participants. Note however that since we only discarded individual trials (rather than participants), this higher exclusion rate should not give rise to confounding individual selection effects (e.g., Bouwmeester et al., 2017).

2.2. Results and discussion

2.2.1. Accuracy

Figure 2A gives an overview of the initial and final accuracies. As the figure indicates, overall, findings are in line with classic results. Participants typically managed to solve incongruent trials correctly when they were allowed to deliberate, although they performed better on congruent than incongruent trials. Regarding initial responses, we overall observed fairly high accuracy rates. For the congruent trials, the mean accuracy for initial responses was 82.6% (SD = 16.7%) and differed from 25% chance ($t(41) = 31.94$, $p < .001$), while for the critical, incongruent trials it was 67.3% (SD = 23.3%) and differed from 25% chance ($t(38) = 18.00$, $p < .001$). This suggests that participants were often able to produce correct responses when deliberation was minimized and they were forced to rely on intuitive, automatic processing. Although this is expected for congruent trials in which the intuitively cued response is correct, it suggests that correct responding on incongruent trials does not necessarily require deliberate controlled processing. To see if there was an effect of the response stage (initial; final) and the congruency status (congruent; incongruent) on the accuracy of the Stroop responses, a two-way within-subjects ANOVA was conducted. As Figure 2A shows, the accuracy for congruent trials was higher than for incongruent trials ($F(1,44) = 15.06$, $p < .001$, $\eta^2 g = 0.048$) and the accuracy at the final stage was higher than at the initial stage ($F(1,44) = 65.83$, $p < .001$, $\eta^2 g = 0.194$), indicating that accuracy improved after deliberation. Finally, the difference between initial and final accuracy was higher for incongruent compared to congruent trials, as indicated by the response stage by congruency interaction ($F(1,44) = 11.08$, $p < .01$, $\eta^2 g = 0.015$).

Note that, in theory, correct responding could result from random guessing. Since our test procedure is highly challenging, participants might not manage to process the stimuli and might respond randomly instead. However, if that were true, accuracy rates should not differ between congruent and incongruent trials and should remain at chance levels throughout the study. It is clear from our findings that this is not the case.

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

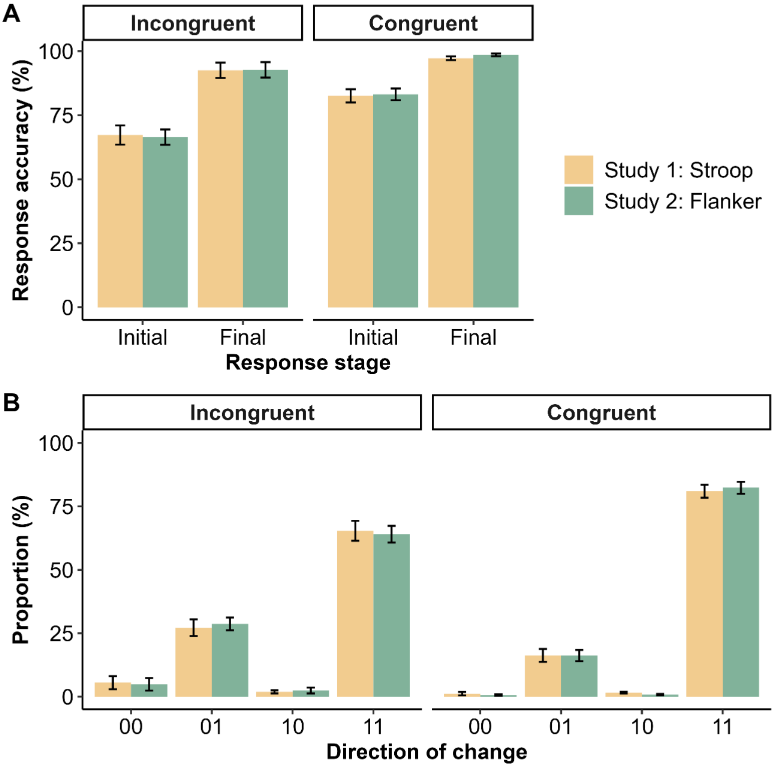

Figure 2. *Accuracy and Direction of change in Study 1 (Stroop task) and Study 2 (Flanker task). (A) Response accuracy at incongruent and congruent trials as a function of the response stage. (B) The proportion of each direction of change category at incongruent and congruent trials. The error bars represent the Standard Error of the Mean. 00, incorrect initial and incorrect final response; 01, incorrect initial and correct final response; 10, correct initial and incorrect final response; 11, correct final and correct initial response.*

In sum, the final accuracy findings are consistent with those of previous Stroop studies (e.g., Aïte et al., 2016). The key finding is the high initial accuracy rate on the incongruent trials. Although accuracy increased in the final stage, we frequently observed correct responding when deliberate control was minimized.

2.2.2. Stability index

We also calculated a stability index for the initial responses of the critical, incongruent trials. Specifically, for each participant, we calculated on how many out of their initial responses in the incongruent trials they showed the same dominant accuracy (i.e., ‘0’ or ‘1’; e.g., if out of 100 trials 60 were incorrect, the stability index would be 60%; similarly, if 60 trials were correct, the stability index would be 60%, etc.). The average stability index was 76.1% (SD = 12.1%). If initial responding was prone to systematic guessing, we would expect more inconsistency in participants’ initial responses across trials.

2.2.3. Direction of change

To get a more precise picture of how participants changed their responses after deliberation, we also conducted a direction of change analysis (Bago and De Neys, 2017, 2019a). More specifically, we looked into how the accuracy changed (or did not change) from the initial to the final stage on every

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

trial. In every stage, participants can either have an accuracy of ‘1’ (i.e., correct response) or an accuracy of ‘0’ (i.e., incorrect response). This way, we end up with 4 possible response patterns in each trial: ‘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).

Regarding the critical incongruent trials, as Figure 2B shows, the vast majority had a ‘11’ pattern (65.4%). This high ‘11’ proportion was also accompanied by a low ‘00’ proportion (5.5%), and a low ‘10’ proportion (1.9%). Critically, the proportion of ‘01’ responses (27.2%) is lower than that of ‘11’ responses. This indicates that, although deliberate correction occurs, in the majority of trials with correct final responses, the correct response was generated already from the initial stage. This so-called non-correction rate (i.e., proportion 11/11 + 01) reached 70.6%.

For completeness, as Figure 2B shows, a similar pattern was observed for congruent trials. In the vast majority of cases, correct responses were generated intuitively. The non-correction rate reached 83%. Again, since intuitive, automatic processing is expected to cue the correct response on these trials, this pattern is not surprising.

2.2.4. Response mapping

A potential difficulty that arises from the specific Stroop task version we adopted is that participants may have struggled to apply the 4-option color-response key mapping during the initial stage. In order to respond, participants first need to identify the color and then translate it into a button press. Despite the time and load constraints during the initial stage, participants likely had enough time to identify the color. However, the complex 4-response mapping may have interfered with translating the color into a button press, which would lead to random guessing. If this was the case, the high accuracy observed in the initial stage could be attributed to guessing.

To examine this further, we looked into the types of errors participants made. Specifically, in the Stroop task, people could make 2 errors: lure errors (responding with the read word instead of the correct ink color) and non-lure errors (responding with any other incorrect ink color). If participants were responding randomly due to time and load constraints, we would expect non-lure errors to occur as frequently as lure errors (i.e., at a chance level of 66.6% and 33.3%, respectively, considering that on each trial participants could make 3 different wrong button presses; 2 non-lure and 1 lure). If, however, participants had sufficient time to press the intended buttons, we would expect primarily lure errors, since participants would be influenced by the read word.

To examine this, we visualized the proportion of lure errors out of all initial errors for each participant separately as a function of initial accuracy (Figure 3A). We conducted a binomial test for each participant’s data, to determine if the lure error proportion exceeded the chance level of 33.3%. In the graph, green dots indicate a significant effect at $p < .05$ (one-tailed), blue dots indicate significance at $p < .01$, and red dots indicate a non-significant effect.

As expected, participants with very high initial accuracy rarely obtained a low p-value since they made very few errors (i.e., they had a low proportion of both lure and non-lure errors). Critically, however, the majority of data points in the upper right corner of Figure 3 are either green or blue. This means that even among participants with high accuracy, the majority of errors were lure errors. This suggests that their high accuracy cannot be attributed to random guessing.

Finally, it is worth noting that a few participants had only non-lure errors. This could result from the fact that they were systematically wrong about the translation of colors into button presses. The existence of these participants indicates that the color-key mapping of the Stroop task was not trivial to learn.

2.2.5. Reaction times

The average reaction time at the initial response stage was 543 ms (SD = 104 ms) for the congruent trials and 581 ms (SD = 55 ms) for the incongruent trials. This is much faster than the average reaction times usually found in previous Stroop studies (e.g., Aïte et al., 2016; Penner et al., 2012; Strauss et al., 2005; Wright and Wanley, 2003) and our one-response control study. Together with the high percentage

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

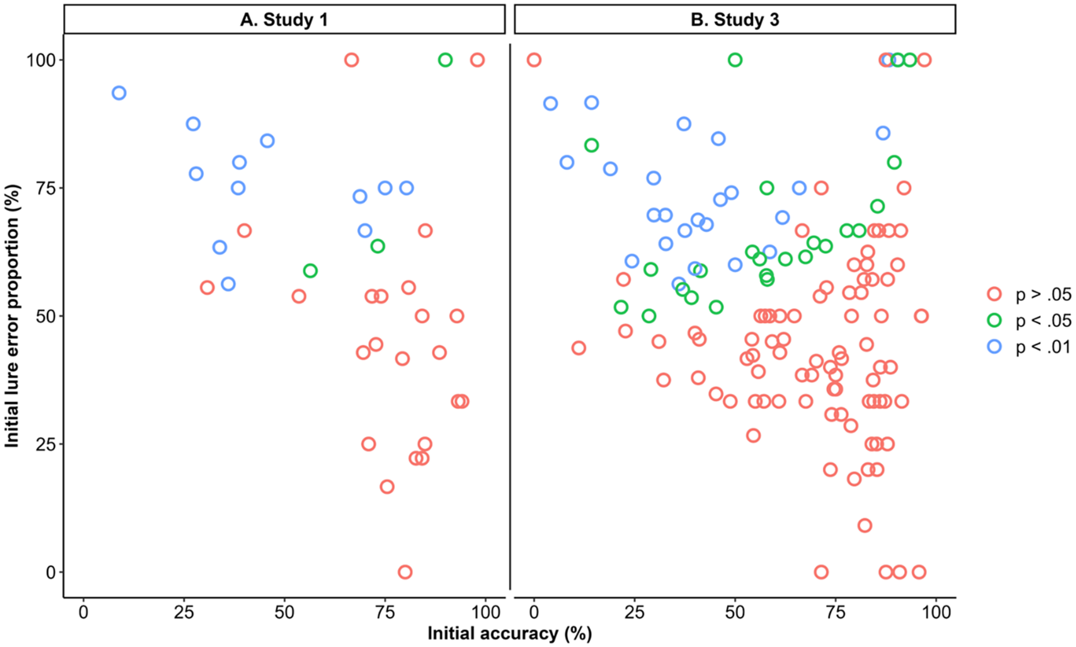

Figure 3. *The initial lure error proportion (% of lure errors out of all errors) as a function of initial response accuracy, separately for each participant in Study 1 (A) and Study 3 (B). A binomial test was conducted for each participant to determine whether the proportion of lure errors exceeded the chance level of 33.3%. Red dots indicate a non-significant effect, green dots indicate a significant effect at $p < .05$ (one-tailed) and blue dots indicate significance at $p < .01$.*

of missed trials, it shows that participants experienced considerable time pressure. Participants spent longer on the final response stage, with an average of 890 ms (SD = 873 ms) for congruent trials and 982 ms (SD = 744 ms) for incongruent trials. Supplementary Material Section B gives a full overview of reaction times according to response accuracy.

2.2.6. Exploratory analysis

To make maximally sure that participants did not deliberate during the initial response stage, we excluded a considerable amount of trials. In theory, this could have artificially boosted the critical non-correction rate. That is, if these excluded trials would be specifically of the ‘01’ type, the true non-correction rate would obviously be lower suggesting that correct intuitive response generation would be much rarer than reported here. To examine this possibility, we re-ran the direction of change analysis while including all missed load and missed deadline trials. Since in the missed deadline trials, the initial response was not recorded, we opted for the strongest possible test and coded all these as ‘0’ (i.e., incorrect response). In the missed load trials both initial and final responses were recorded. The analysis (see Supplementary Material Section C for full results) pointed to a higher proportion of ‘01’ incongruent trials (47.2%), but the proportion of ‘11’ (41.2%) responses and the non-correction rate remained high (46.6%). Hence, even in this extremely conservative analysis, correct incongruent responses were still generated intuitively about half of the time.

3. Study 2

Study 1 showed that when participants gave a correct final Stroop response they had typically already generated a correct response in the initial stage. This indicates that correct responding in the Stroop task can occur even when deliberate control is minimized. However, Study 1 was but the first to

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

adopt the two-response paradigm with a classic cognitive control task. Thus, it is important to test the generalizability of these findings to another classic cognitive control task before drawing strong conclusions. Therefore, in Study 2, we designed a two-response version of the Flanker task. Since the Flanker task is a binary-response task, it also allowed us to sidestep the difficulty of the specific Stroop task response format we adopted in Study 1, namely that participants may have found it challenging to apply the 4-option color-response key mapping in the initial stage.[2] As in Study 1, the key question is whether participants can provide correct responses to the critical incongruent Flanker trials when their deliberate control is constrained.

3.1. Method

3.1.1. Preregistration and data availability

The study design and hypothesis were preregistered on the Open Science Framework (https://osf.io/eqdks). No specific analyses were preregistered. All data are also available on the Open Science Framework (https://osf.io/gkhbm/).

3.1.2. 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. Participants were paid £2.40 for their participation (£6 hourly rate).[3] For consistency with Study 1, we recruited 50 adult participants. The mean age of participants was 38.6 years (SD = 14.6) and 58% were female. Thirty-eight percent of participants had a high-school degree as their highest education level, 46% had a bachelor’s degree, 12% had a Master’s degree, 2% had a doctoral degree, and 2% had not completed high school.

3.1.3. Materials

3.1.3.1. Flanker stimuli

The stimuli consisted of a row of 5 arrows. This row included a central arrow flanked by 2 surrounding arrows on each side, all with arrowheads pointing either to the left or to the right. In congruent stimuli, the surrounding arrows pointed in the same direction as the central arrow (←←←←← or →→→→→). In incongruent stimuli, the surrounding arrows pointed in the opposite direction to the central arrow (←←→→←← or →→←→→).

A total of 128 experimental trials, consisting of 64 congruent and 64 incongruent trials, were presented to the participants in a randomized order. The stimuli were presented in the center of the screen on a white background. Participants were instructed to press the ‘f’ key if the central arrow pointed left and the ‘j’ key if it pointed right. Response times were measured from the onset of the stimulus until the button press. Our main results concern the critical incongruent trials unless otherwise stated.

As we noted, since the Flanker task is a binary-response task, it also allows us to sidestep a potential difficulty of the specific Stroop task response format we adopted in Study 1, namely that participants may have found it challenging to apply the 4-option color-response key mapping in the initial stage. However, in theory, the version of the Flanker task that we used may present its own limitations. For example, one may note that in the congruent trials, it is not necessary to focus attention on the central arrow, since all items are identical, but in the incongruent trials participants need to focus their attention on the central arrow to produce a correct response. This may invite an alternative strategy that people

[2] However, note that although both the Flanker task and the Stroop task involve conflict resolution in the incongruent trials, they tap into different aspects of cognitive control, and while the Stroop involves semantic conflict, the Flanker involves a more perceptual conflict (Ridderinkhof et al., 2021; Sections 3.1 and 5).

[3] The hourly rate in this study is £6 instead of the £5 hourly rate of Studies 1 and 3, as Prolific increased their minimum pay by the time Study 2 was run.

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

can use: they can first determine whether all items are the same and, if they are not, they can focus their attention on the central target only. Since focusing takes time this strategy could generate longer reaction times in the incongruent, compared to the congruent trials. In this sense, the Flanker task would not necessarily evoke response conflict like the Stroop. However, the evidence in the cognitive control literature with the specific version of the Flanker task (with a 1-to-1 response mapping) we adopted suggests that this alternative account is insufficient to explain the entirety of the flanker effect (e.g., Hübner et al., 2010) and may not even play a significant role in contributing to it (Servant and Logan, 2019). That is because participants focus attention on the central arrow in a similar way in congruent and incongruent trials (Servant and Logan, 2019). This supports the original interpretation of the Flanker, which emphasizes response competition as a key factor in the task (Eriksen and Eriksen, 1974; Eriksen and Hoffman, 1973). Nevertheless, it remains the case that the Stroop and Flanker tasks may tap different aspects of cognitive control (e.g., Friedman and Miyake, 2004; Rey-Mermet et al., 2018; Section 5).

3.1.3.2. Load task

In the two-response version of the Flanker task, we used the same secondary digit memorization task as in the Stroop task of Study 1 (Lavie, 2005; Lavie et al., 2004), since it has been shown to burden cognitive control in classic control tasks (Lavie et al., 2004).

3.1.4. Procedure

3.1.4.1. One-response (deliberative-only) pre-test

To obtain a baseline Flanker performance, we ran a pre-test where participants performed a traditional one-response arrow Flanker task, without a digit memorization load or a deadline. As in Study 1, we recruited 25 participants (48% female; mean age = 36.4 years, SD = 11.0) 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. Participants were paid £0.70 for their participation. A total of 40% of the participants reported a bachelor’s degree as their highest education level, while 28% reported a Master’s degree and 28% a high school degree.

The deadline of the initial response stage in our two-response design was based on the average response time of the one-response pretest (e.g., Bago and De Neys, 2017, 2020). Thus, the one-response pretest was similar to the main study in terms of stimuli and amount of trials, but participants were instructed to provide a single response on each trial and to answer ‘as fast and as accurate as possible’. The average response time for the congruent trials was 428 ms (SD = 52.2 ms) and for the incongruent trials, it was 458 ms (SD = 47.2 ms).[4] Based on these values we decided to set the maximum response deadline for the initial response to 420 ms (i.e., approximately the mean of congruent trials which do not require controlled processing to answer correctly).

To confirm that participants were under time pressure in the initial stage, we compared response times for critical incongruent trials between the one-response pre-test and the initial responses in the main two-response study. We first excluded all trials with missed load memorization or missed deadlines in the initial stage of the two-response study. The results showed that participants responded much faster in the initial response stage of the main study (incongruent trials: 314.6 ms, SD = 43.7 ms), compared to that of the one-response study (incongruent trials: 457.6 ms, SD = 47.2 ms; i.e., responses were on average more than 2.5 SDs faster than in the one-response study). A Welch Two Sample *t*-test indicated that this difference was significant ($t(3222.88) = 55.80$, $p < .001$).

The one-response pre-test also allowed us to check for a potential consistency confound in our main two-response study, which could potentially underestimate the correction rate from initial to final responses. To test for this confound, we compared the accuracy of incongruent trials between

[4] Before computing the average reaction times all trials with reaction times higher than 2 SDs above the general mean were removed from the analysis.

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

the final two-response stage of our main study (M = 92.7%, SD = 20.6%) and the pretest (M = 97.5%, SD = 2.5%). Although a Welch Two Sample *t*-test revealed a significant difference (*t*(2792.13) = 5.77, *p* < .001), this difference was small. Even if we factor in a possible 5% extra correction trials (i.e., ‘01’ trials) in our results, the non-correction rate conclusions remain unaffected (i.e., 65.5% with the extra correction trials vs. 69% without). Therefore, a potential consistency confound cannot explain the low correction rates.

3.1.4.2. Two-response Flanker task

The experiment was run online on Gorilla Experiment Builder (gorilla.sc). Participants were informed that the study would take 20 minutes and that it required their full attention. They were told that they would be presented with an arrow at the center of the screen and that they had to press the button that matched the arrow’s direction. Specific instructions about the key mapping were provided. Participants were then told that the central arrow would always appear along with 4 other arrows and that their task was to identify the direction of the central arrow (for literal instructions, see Supplementary Material Section A).

To familiarize themselves with the key mappings, participants first practiced with 6 trials (3 congruent and 3 incongruent). They were given feedback after each response and in case of an incorrect answer they were reminded of the correct key pairs.

At the end of this practice round, participants were introduced to the two-response paradigm. They were told that we were first interested in their initial, intuitive response to the direction of the central arrow and wanted them to answer as fast as possible with the first response that came to mind. They were told that after this first response, they would have more time to reflect before providing their final answer. Participants were introduced to the deadline of the initial response and were shown an example of an initial trial. Then, they were presented with 6 two-response trials.

Following the two-response practice round, participants were presented with the load task, with the same instructions as in Study 1. They were then presented with 5 load memorization practice trials.

After the load practice round, participants were reminded that they had to memorize the numbers while responding to the direction of the central arrow. They were instructed to first focus on the memorization task, and then on the arrow task. They were then presented with 12 two-response practice trials (with load and deadline). Critically, the first 6 practice trials had a looser initial response deadline (670 ms instead of 420 ms). This was done to familiarize participants with the two-response format. For the last 6 practice trials the actual 420 ms deadline was applied.

After this practice session, participants started the experimental trials. The main task was composed of 128 trials which were grouped into 3 blocks. Participants were told that after each block they could take a short break. Before each new block started, they were reminded of the response key mapping. At the end of the experiment, they completed standard demographic questions and were presented with a debriefing message.

3.1.5. Exclusion criteria

Like in Study 1 and following our preregistration, we discarded from all analyses participants who scored lower than 50% on both their initial congruent and initial incongruent trials. Based on this, 1 out of the 50 participants was excluded. We were thus left with a sample of 49 participants (57% female) with a mean age of 38.6 years (SD = 14.6).

In addition, we excluded the trials in which participants failed the load and/or the deadline. Participants failed to answer before the deadline on 39.0% of incongruent initial trials (1222 out of 3136) and 29.3% of congruent initial trials (919 out of 3136). In addition, participants failed the load task on 8.2% of incongruent initial trials (257 out of 3136) and 12.1% of congruent initial trials (381 out of 3136). Overall, we kept 55.7% (3493 out of 6272), by rejecting trials in which participants missed the deadline and failed the load task. On average, each participant contributed 71.3 trials (out of 128 trials, SD = 32.0). As in Study 1, the high number of missed trials indicates that meeting the deadline and load constraints was challenging for participants.

3.2. *Results and discussion*

3.2.1. Accuracy

Figure 2A gives an overview of the initial and final accuracies. Overall the results are very similar to that of the Stroop task of Study 1. Participants generally performed better on congruent trials, but they also managed to solve most incongruent trials correctly when deliberate processing was allowed. Initial responses showed high accuracy rates both for congruent (M = 83.2%, SD = 15.8%) and critical incongruent trials (M = 66.5%, SD = 20.5%) and they both differed from 50% chance ($t(47) = 14.58$, $p < .001$ and $t(45) = 5.45$, $p < .001$, respectively). This suggests that participants often produced correct responses even when relying on mere intuitive processing. So, as in the Stroop task, correct responding in the Flanker task does not necessarily require deliberate controlled processing. A two-way within-subjects ANOVA on the effect of response stage and congruency status on response accuracy revealed that accuracy was higher for congruent trials ($F(1{,}45) = 22.21$, $p < .001$, $\eta^2 g = 0.10$) and that accuracy at the final stage was higher than that at the initial stage ($F(1{,}45) = 100.38$, $p < .001$, $\eta^2 g = 0.29$). This difference between initial and final accuracy was higher for incongruent compared to congruent trials, as indicated by the response stage by congruency interaction ($F(1{,}45) = 10.70$, $p < .01$, $\eta^2 g = 0.02$).

In sum, these results align with the Stroop results of Study 1 and show that correct responding in the incongruent trials of the Flanker task is possible when deliberate control is minimized.

3.2.2. Stability index

We also calculated a stability index for the initial responses of the critical, incongruent trials. More specifically, for each participant, we again calculated how many out of their initial responses in the incongruent trials they showed the same dominant accuracy (i.e., ‘0’ or ‘1’). The average stability index was 71.8% (SD = 14.5%). If initial responding was prone to systematic guessing, we would expect more inconsistency in participants’ initial responses across trials.

3.2.3. Direction of change

To get a more precise picture of how participants changed their responses after deliberation, we again conducted a direction of change analysis (Bago and De Neys, 2017, 2019a). Regarding the critical incongruent trials, as Figure 2B shows, the vast majority had a ‘11’ pattern (64.0%). This high ‘11’ proportion was also accompanied by a low ‘00’ proportion (4.9%), and a low ‘10’ proportion (2.4%). Critically, the proportion of ‘01’ responses (28.7%) was lower than that of ‘11’ responses. This indicates that, although deliberate correction occurs, in the majority of trials with correct final responses the correct response was generated already from the initial stage. The non-correction rate (i.e., proportion 11/11 + 01) reached 69%. As it was expected and as Figure 2B shows, in the vast majority of congruent trials correct responses were generated intuitively and the non-correction rate reached 83.6%.

3.2.4. Reaction times

The average reaction time at the initial response stage was 305.6 ms (SD = 58.3 ms) for the congruent trials and 314.6 ms (SD = 43.7 ms) for the incongruent trials. This is much faster than the average reaction times found in previous Flanker studies with similar amount of trials (e.g., Abutalebi et al., 2012; Fan et al., 2005) and our one-response control study. Together with the high percentage of missed trials, it shows that participants experienced considerable time pressure. Participants spent longer on the final response stage, with an average of 515.2 ms (SD = 225.1 ms) for congruent trials and 543.2 ms (SD = 260.2 ms) for incongruent trials. Supplementary Material Section B gives a full overview of reaction times according to response accuracy.

3.2.5. Exploratory analysis

To ensure that participants did not deliberate during the initial response stage, we excluded a considerable amount of trials, which could have potentially inflated the non-correction rate. To examine this possibility, we re-ran the direction of change analysis while including all missed load and missed deadline trials. As in Study 1, we opted for the strongest possible test and coded all missed deadline trials as ‘0’ (i.e., incorrect response). In the missed load trials both initial and final responses were recorded. The analysis (see Supplementary Material Section C for full results) pointed to a higher proportion of ‘01’ incongruent trials (52.3%), but the proportion of ‘11’ (39.2%) responses and the non-correction rate remained high (42.8%). Hence, even in this extremely conservative analysis, correct incongruent responses were still generated intuitively about 43% of the time.

4. Study 3

Studies 1 and 2 showed that in both the Stroop and Flanker tasks, when participants provided a correct final response, they had typically already generated a correct response in the initial intuitive stage. This indicates that correct responding in cognitive control tasks is possible even when deliberate control is minimized. The first aim of Study 3 was to replicate the Stroop findings of Study 1 on a larger scale. The second aim was to explore whether individual performance in the Stroop task, both at the initial and final stage, correlates with performance in classic heuristics-and-biases tasks.

Study 3 comprised two parts: a Color-Word Stroop task followed by a Reasoning task consisting of a battery of heuristics-and-biases reasoning problems. We used a two-response paradigm (Thompson et al., 2011) for both the Stroop and the Reasoning task.

4.1. Method

4.1.1. Preregistration and data availability

The study design and hypothesis were preregistered on the Open Science Framework (https://osf.io/dm7h9). No specific analyses were preregistered. All data and material are also available on the Open Science Framework (https://osf.io/yqkm7/).

4.1.2. 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. Participants were paid £4.50 for their participation (£5 hourly rate). Based on Raoelison et al. (2020), Study 2) correlational two-response study, we aimed to recruit 160 participants. Due to a software error, the Reasoning task data of one participant could not be recovered, so we ended up with 159 participants (69.2% female), with a mean age of 33.1 years (SD = 13.6). This allowed us to pick up small to medium-size correlations (.22) between the Stroop and Reasoning task performance with a power of 80%. The majority of participants (45%) had a high-school degree as their highest education level, 37% had a bachelor’s degree, 15% had a Master’s degree, and 3% had not completed high school.

4.1.3. Materials

The Stroop task was run on Gorilla Experiment Builder (gorilla.sc) and the Reasoning task was run on the Qualtrics (www.qualtrics.com) software server. We first ran an initial batch of 10 participants that was identical to the main study. This was done to ensure that no technical problems would occur during the transition from Gorilla Experiment Builder to the Qualtrics platform. Data from one participant of this first batch could not be analyzed (see above). We then ran the main study batch, which consisted of the remaining 150 participants.

The Color-Word Stroop task that was used in this study was identical to the Stroop task described in Study 1.

The Reasoning task included 3 different types of reasoning problems (i.e., bat-and-ball problems, base-rate problems, and syllogistic reasoning problems). We used the exact same two-response format (response deadlines and load, see below) that was validated for these tasks in previous work (Bago and De Neys, 2017, 2019a; De Neys, 2006). To avoid confusion, it is important to stress that the deadline and the concurrent cognitive load of the Reasoning task differs from that of the Stroop task. As a reminder, the goal of these 2 constraints is to minimize deliberation involvement and enforce intuitive thinking. However, there is no gold standard procedure which can ensure that people will respond intuitively, and the definition of 'limited cognitive resources' always depends on the task at hand. For example, heuristics-and-biases tasks are lengthy (e.g., a couple of preamble sentences and response option reading), so deadlines are based on the pretested average reading times which are usually a couple of seconds (participants need to have the minimum time to read the problem before responding). On the contrary, Stroop responding is considerably faster since participants only see a single stimulus (i.e., word), so a strict deadline necessarily cannot be much longer than a single second. The same goes for the cognitive load, whose goal is to burden participants' cognitive resources. The strain on resources may depend on the specific nature of the task. That is why, for each of our tasks, we opted for a load that has been independently shown in the literature to burden cognitive resources and decrease performance in this specific type of task.

4.1.3.1. Counterbalancing

Each of the 3 types of reasoning problems was composed of 8 incongruent and 8 congruent items. For every type of problem, we created 2 sets of items. In each set, the congruency status of the items was counterbalanced. More specifically, all the incongruent items of the first set appeared in their congruent version in the second set, and all the congruent items in the first set appeared in their incongruent version in the second set. Half of the participants were presented with the first set while the other half were presented with the second set. This way, the same item content was never presented more than once to a participant and, at the same time, everyone was exposed to the same items, which minimized the possibility that mere item differences influence the results (e.g., Bago and De Neys, 2017). The presentation order of the items within each task was randomized. Each participant was randomly allocated to one of 6 potential task orders. More specifically, each participant was first randomly allocated to task 1/3 (i.e., either bat-and-ball, base-rates, or syllogisms), and then they were randomly allocated to one of the 2 potential task order combinations for the second and third task (e.g., if a given participant had the bat-and-ball as their first task, they could continue with base-rates as their second task and syllogisms as their final task, or the inverse).

4.1.3.2. Bat-and-ball problems (BB)

Each participant was presented with 8 multiple-choice bat-and-ball items (4 incongruent and 4 congruent) taken from Bago and De Neys (2019a). The prices and the names of the objects varied between items, but all the items shared the same structure with the classic bat-and-ball problem. Participants were always presented with 4 response options: the logical option ('5 cents' in the original bat-and-ball), which is considered correct, the heuristic option ('10 cents' in the original bat-and-ball), and 2 foil options. The 2 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 divisor (e.g., '1 cent' in the original). An example of the problems is presented below:

> A pencil and an eraser cost $1.10 in total.
>
> The pencil costs $1 more than the eraser.
>
> How much does the eraser cost?

○ 5 cents
○ 1 cent
○ 10 cents
○ 15 cents

The congruent versions were constructed by removing the 'more than' statement from the incongruent versions ('A pencil and an eraser cost $1.10 in total. The pencil costs $1. How much does the eraser cost?'). Each problem was presented serially. First, the first sentence, which always stated the 2

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

objects and their total cost (e.g., A pencil and an eraser cost $1.10 in total.) was presented for 2000 ms. Afterward, the second sentence along with the question and the answer options was added under the first sentence (which remained on screen). The problem remained on the screen until a response was given or until the deadline. As in Bago and De Neys (2019a), the deadline for the initial response was 5000 ms.[5]

4.1.3.3. Base-rate problems (BR)

Each participant was presented with 8 base-rate items (4 incongruent and 4 congruent) taken from Bago and De Neys (2017). Each item consisted of a sentence describing the composition of a sample (e.g., ‘This study contains scientists and assistants’.), a sentence with a stereotypical description of a random person from the sample (e.g., ‘Person “C” is intelligent’.), and a sentence with the base-rate information (e.g., ‘There are 4 scientists and 996 assistants’.). Participants had to indicate to which group the random person most likely belonged to. The answer option that was considered correct was always the one that corresponded to the largest group in the sample. The presentation of all items was based on Pennycook et al.’s (2014) rapid-response paradigm. Each sentence was presented serially and the amount of text presented on the screen was minimized. An example of the problems is presented below:

This study contains scientists and assistants.

Person “C” is intelligent.

There are 4 scientists and 996 assistants.

Is Person “C” more likely to be:

○ A scientist
○ An assistant

The congruent versions were constructed by reversing the base rates of the incongruent versions. For example in its congruent version, the second sentence of the above problem would read ‘There are 996 scientists and 4 assistants’. Each problem was presented in 3 stages. First, the first sentence was presented for 2000 ms. Then, the second sentence was added under the first sentence (which remained on screen) for another 2000 ms. 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.

4.1.3.4. Syllogistic reasoning problems (SYL)

Each participant was presented with 8 syllogistic reasoning items (4 incongruent and 4 congruent), taken from Bago and De Neys (2017). Each item consisted of a major premise (e.g., ‘All things made of wood can be used as fuel’.), a minor premise (e.g., ‘Trees can be used as fuel’.) and a conclusion (e.g., ‘Trees are made of wood’.). 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 things made of wood can be used as fuel

Trees can be used as fuel

Trees are made of wood

Does the conclusion follow logically?

○ Yes
○ No

[5] The specific deadlines in each type of problem were based on pilot reading and one-response pretests (see respective subsections) and have been shown to create substantial time pressure.

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

In the incongruent items, the believability and the validity of the conclusion conflicted. More specifically, the conclusion of the incongruent items was either valid-unbelievable or invalid-believable. For instance, in the above example of an incongruent problem the syllogism is believable, but invalid. For the congruent items, the validity of their conclusion was in accordance with their believability. Meaning that the conclusion was either valid-believable or invalid-unbelievable. For example, in its congruent version, with a valid-believable conclusion, the above problem would read: 'All things made of wood can be used as fuel. Trees are made of wood. Trees can be used as fuel'. Each problem was presented in 3 stages. First, the first sentence of the problem was presented for 2000 ms. Then, the second sentence was added under the first sentence (which remained on screen) for 2000 ms. 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.

4.1.3.5. Load task

For the Stroop task, we used the same digit memorization task (Lavie, 2005; Lavie et al., 2004) as in Study 1. For the Reasoning task, the load memorization task that was used was a complex visual pattern (i.e., 4 crosses in a 3 × 3 grid, Bago and De Neys, 2017, 2019a; Raoelison and De Neys, 2019), which was briefly presented before each reasoning problem (Miyake et al., 2001). After providing an initial response to the reasoning problem, participants were presented with 4 different load patterns (i.e., with different cross placings) and had to identify the one that they had been asked to memorize. Miyake et al. (2001) showed that this task burdens cognitive resources, and previous studies have shown that it hampers sound deliberating and decreases reasoning accuracy on the specific types of reasoning problems we adopted (e.g., De Neys, 2006; Franssens and De Neys, 2009; Johnson et al., 2016).

4.1.3.6. Composite reasoning measure

For simplicity and to maximize power, our analyses focused on the composite incongruent accuracy across the 3 different reasoning problem types (i.e., bat-and-ball, base rates, syllogisms). To calculate the composite performance, we averaged for each participant the proportion of correct initial and final responses, separately for each problem type. Then we averaged across all problem types (separately for initial and final trials). For completeness, we calculated the composite performance also for congruent trials.

For the main correlational analysis between the Stroop and the Reasoning task, we first calculated the *z*-scores separately for each participant, each problem type, each response stage (i.e., initial, final), and each direction of change category (see further). Then, we averaged the *z*-scores across the 3 problem types, separately for each response stage and each direction of change category.

It is important to clarify that because of practical limitations, we did not have a composite cognitive control measure. Thus, we tested whether the composite reasoning measure correlated with performance at the Stroop task only.

4.1.4. Procedure

Participants were informed that the study would take 55 minutes to complete and that it demanded their full attention. They were told that the experiment was divided into 2 parts (i.e., the Stroop task and the Reasoning task). All participants began the experiment with the Stroop task, and once they finished, they were redirected to the Reasoning task. The Stroop task's procedure was identical to the one described in Study 1. Once participants started the Reasoning task, they were told that it consisted of 3 different types of reasoning problems (i.e., bat-and-ball, base rates, and syllogisms). Then, they were told that they would have to provide 2 consecutive responses to various items. They were instructed 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 Raoelison et al., 2020, for literal instructions).

Afterward, participants were presented with instructions specific to each problem type. Each problem type made up a block of the task and the 3 different types were presented in a pseudorandomized order (Section 4.1.3.1). Every problem type was introduced with a short transition text which indicated

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

the participant's progress (e.g., 'You are going to start task 1/3. Click on Next when you are ready to start task 1'.). Then, the presentation format of the respective problem type was explained, an example problem was shown, and the deadline of the initial response was introduced. After these instructions, participants solved 2 practice items (without a concurrent load task) to familiarize themselves with the presentation format. Next, they solved 2 practice matrix recall items (without a concurrent reasoning problem). Finally, they solved the 2 earlier practice items with a concurrent load task.

Each trial started with a fixation cross that was shown for 1000 ms. Next, the target pattern for the memorization task was presented for 2000 ms. Then the first part of the problem was presented (for more details see Materials subsections for each problem type). Afterward, 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 of the upcoming deadline. 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).[6] After entering their confidence, participants were shown 4 matrix patterns and were asked to recall the correct, to-be-memorized pattern. They were then given feedback on whether their recall was correct or not. Finally, participants viewed the full problem again and were asked to provide their final answer. Next, they were asked to report their confidence in the correctness of their final response. After responding to all the items of a problem type, a transition message appeared to indicate participants' progress (e.g., 'You are going to start task 2/3. Click on Next when you are ready to start task 2'.). At this point, the next problem type was introduced.

After participants had responded to all 3 problem types, they were shown the classic bat-and-ball problem and were asked whether they had seen or read about this specific problem before (Yes/No). Immediately afterward they were asked to provide an answer to the problem ('What do you think the correct answer is? Please enter it below'). Finally, participants were asked to complete standard demographic questions and were shown a debriefing message.

4.1.5. Exclusion criteria

As in Study 1, we discarded from all analyses participants who scored lower than 50% on both their initial congruent and initial incongruent Stroop trials. As a result, 13 out of the 159 participants were excluded. We were thus left with a sample of 146 participants (59% female), with a mean age of 36.6 years (SD = 14.1).

4.1.5.1. Stroop task

Participants did not respond within the deadline on 18.1% of congruent initial trials (1690 out of 9344) and 29.2% of incongruent initial trials (2728 out of 9344). In addition, participants failed the load recall on 15.2% of congruent initial trials (1416 out of 9344) and 10.5% of incongruent initial trials (979 out of 9344). By rejecting the trials with a missed deadline and an incorrect load recall, we kept 63.5% of all trials (11875 out of 18688). On average, each participant contributed 81.3 trials (out of 128 trials, SD = 31.5).

4.1.5.2. Reasoning task

The trials in which participants failed the load and/or the deadline were excluded from subsequent analyses. Participants failed to answer before the deadline on 5.4% of incongruent initial trials (103 out of 1908) and 2.7% of congruent initial trials (52 out of 1908). In addition, participants failed the load recall on 12.5% of incongruent initial trials (239 out of 1908) and 14.7% of congruent initial trials (281 out of 1908). By rejecting the trials with a missed deadline and an incorrect load recall, we kept 82.3%

[6] The confidence was recorded both at the initial and the final responses simply for a comparison with previous reasoning findings (Supplementary Material Section F).

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

of all trials (3141 out of 3816). On average, each participant contributed 19.8 trials (out of 24 trials, SD = 3.1).

Since the bat-and-ball problem has become very popular, some participants may have been previously exposed to the correct ‘5 cents’ answer. If this is the case, they would not need to override an initially incorrect, heuristic response in order to arrive at the correct answer when solving the problem, which could distort our results. Following Raoelison et al. (2020), we, therefore, asked participants whether they had seen/solved the bat-and-ball problem before or if they had read about it (Section 4.1.4). We also asked them to provide an answer to the problem (‘What do you think the correct response is? Please enter it below’.). The bat-and-ball trials were excluded for all participants that reported having seen the original bat-and-ball problem and that were able to provide the correct ‘5 cents’ response.[7] Their trials for the other tasks were included in the analysis. In total, we excluded from the analysis an additional 440 bat-and-ball trials (i.e., 5.8% of all trials) from 32 participants. Note that, 56 of the bat-and-ball trials of these participants were already excluded because of missed deadline or load.

4.2. *Results and discussion*

4.2.1. Stroop task

In Study 3, we replicated the main findings observed in the Stroop task of Study 1, with a much larger sample. Specifically, we found that participants can typically provide correct Stroop responses, even when deliberate control is minimized. The mean accuracy at congruent trials was 63.6% (SD = 24.3%) in the initial response stage and 91.5% (SD = 17.9%) in the final stage, while the non-correction rate (i.e., proportion 11/11 + 01) reached 66.5%. These results again suggest that, more often than not, correct Stroop responses are generated in the absence of deliberate controlled correction. For brevity, the full results of the Stroop task of Study 3 are reported in Supplementary Material Section D.

4.2.2. Reasoning task

4.2.2.1. Accuracy

Figure 4 gives an overview of the initial and final Reasoning task accuracies. Although we focus our analysis on the composite reasoning performance, individual task trends are reported in the graphs for completeness. The overall pattern is very similar to what was observed in previous two-response studies. First, people perform well on congruent trials both at the initial (M = 90.0%, SD = 7.2%) and the final stage (M = 92.6%, SD = 7.1%), while incongruent trials typically have low initial (M = 35.2%, SD = 20.5%) and final (M = 40.6%, SD = 22.6%) accuracies. This indicates that even after deliberation, the majority of reasoners remain biased (Bago and De Neys, 2017, 2019a; Raoelison et al., 2020; Raoelison and De Neys, 2019). As it can be seen in Figure 4, these composite level trends were also observed for each individual task separately.

In addition, note that consistent with previous findings, reasoners’ accuracy at the incongruent trials is typically below or near 50%, (and close to guessing accuracy). However, the high accuracy on the congruent trials confirms that participants are not merely guessing throughout the study. Instead, they are simply lured by the heuristic cue when solving the incongruent items.

To examine whether there was an effect of the response stage (initial; final) and congruency status (incongruent; congruent) on response accuracy, a two-way within-subjects ANOVA was conducted. As Figure 4 shows, the accuracy at the congruent trials was higher than that at the incongruent trials ($F(1,158) = 402.54$, $p < .001$, $\eta^2 g = 0.518$), and the accuracy at the final stage was higher than that at the initial stage ($F(1,158) = 29.91$, $p < .001$, $\eta^2 g = 0.008$), showing that accuracy improved after deliberation. Finally, this difference between initial and final accuracy was higher for

[7] The answer to this question was in free response format. The responses that were considered as correct were: 5 cents, 5 CENTS, 5c, 5, $0.05, 0.05, .05, and 0.5.

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

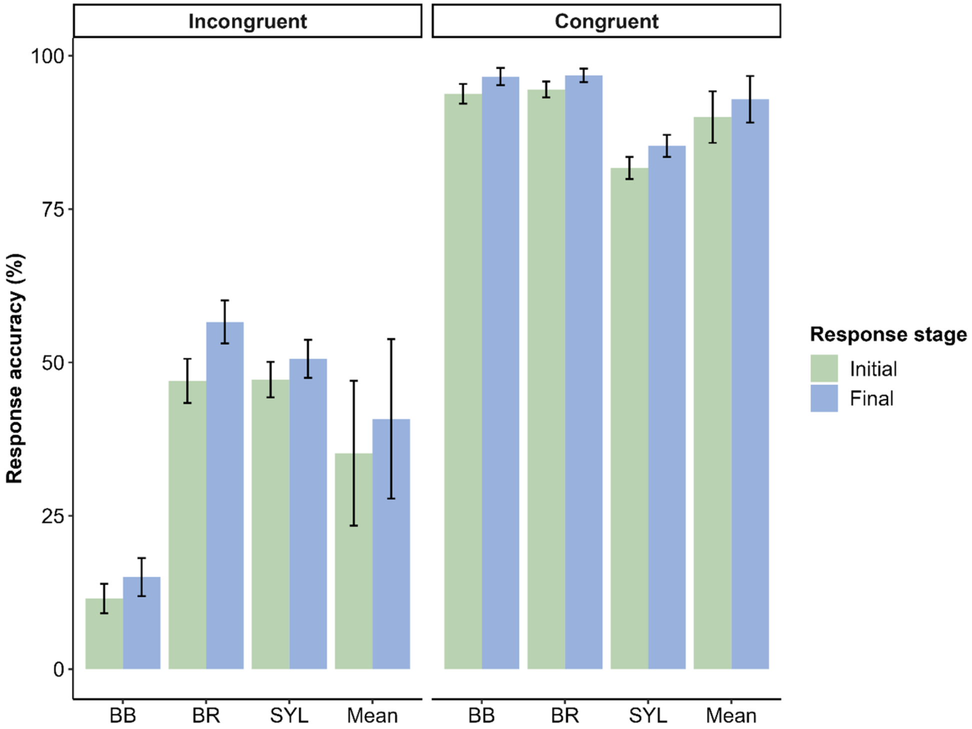

Figure 4. *Response accuracy at incongruent and congruent trials of the Reasoning task in Study 3 for initial and final responses, separately for each problem type and for the mean across the 3 problem types. The error bars represent the Standard Error of the Mean. BB, bat-and-ball; BR, base-rates; SYL, syllogisms; Mean, the mean across the 4 problem types.*

incongruent compared to congruent trials, as indicated by the response stage by congruency interaction ($F(1,158) = 4.08, p < .05, \eta^2 g = 0.001$).

4.2.2.2. Stability index

Like in the Stroop task, we calculated a stability index for the initial responses of the critical, incongruent trials. For each participant, we calculated on how many out of their initial responses in the incongruent trials they showed the same accuracy (i.e., ‘0’ or ‘1’). The average stability index was 95.5% (SD = 11.1%) in the bat-and-ball task, 93.0% (SD = 14.6%) in the base-rate task, and 81.3% (SD = 19.6%) in the syllogistic reasoning task. If initial responses were susceptible to systematic random guessing, we would observe more inconsistency in response patterns.

4.2.2.3. Direction of change

To get a more precise picture of how participants changed their responses after deliberation we also conducted a direction of change analysis (Bago and De Neys, 2017, 2019a). As Figure 5A shows, at the composite level, the majority of the critical, incongruent trials had a ‘00’ pattern (52.2%) which confirms that reasoners are easily lured by the heuristic response when solving reasoning items. Critically, in the incongruent trials, the proportion of ‘11’ responses (35.2%) is higher than that of the ‘01’ responses (9.2%). The mean composite non-correction rate (i.e., proportion 11/11 + 01) reached 79.3%. Hence, as in the Stroop task, although there is some accuracy increase after deliberation, correct responses are, for the most part, already generated intuitively.

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

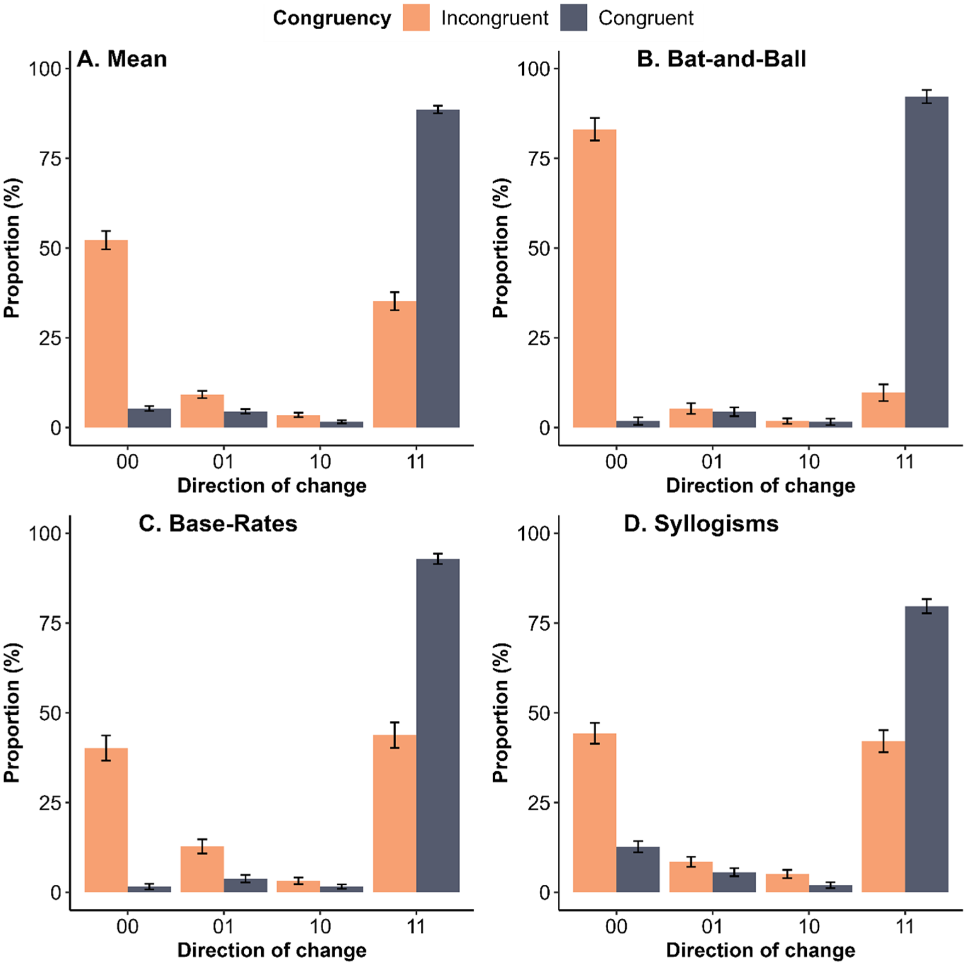

Figure 5. *Direction of change in the Reasoning task of Study 3 separately for incongruent and congruent trials. (A) The proportion of each direction of change category at the composite level. (B) The proportion of each direction of change category at the Bat-and-Ball trials. (C) The proportion of each direction of change category at the Base-Rates trials. (D) The proportion of each direction of change category at the Syllogistic reasoning trials. The error bars represent the Standard Error of the Mean. 00, incorrect initial and incorrect final response; 01, incorrect initial and correct final response; 10, correct initial and incorrect final response; 11, correct final and correct initial response.*

4.2.3. Correlation between Stroop and reasoning

We now turn to the main analysis of Study 3 in which we explore the relationship between participants' performance on the Stroop task and Reasoning task. For simplicity, we always use the Stroop task as the predictor of the Reasoning performance when interpreting the results.

For completeness, we also computed the split-half reliability of incongruent trials in both the Stroop task and the Reasoning task, separately for initial and final responses. The split-half reliability in the Stroop task was 0.94 for initial responses and 0.97 for final responses. In the bat-and-ball task, the split-half reliability was 0.78 for initial and 0.98 for final responses, in the base-rate task it was 0.91 for initial and 0.89 for final responses, and in the syllogistic reasoning task, it was 0.66 for initial and 0.65 for final responses. For the composite reasoning measure, the split-half reliability was 0.82 for initial and 0.86 for final responses.

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

Table 1. *Pearson's product–moment correlation between the average accuracy of each individual on the Stroop task, and the accuracy of that individual on the Reasoning task.*

Reasoning accuracy		Stroop accuracy			
		Initial		Final	
		r	*p*	*r*	*p*
Initial	BB	−0.05	0.584	0.14	0.142
	BR	0.07	0.426	0.14	0.092
	SYL	0.08	0.343	0.09	0.277
	Composite	−0.06	0.569	0.14	0.145
Final	BB	−0.09	0.327	0.14	0.154
	BR	0.04	0.674	**0.18**	0.037
	SYL	0.07	0.432	0.10	0.266
	Composite	−0.12	0.234	0.16	0.109

Note: Correlations are reported both at the composite level and for each type of reasoning problem, separately for each Response stage (initial response; final response). Significant correlations ($p < .05$) are in bold.
BB, bat-and-ball; BR, base-rates; SYL, Syllogisms.

4.2.3.1. Accuracy

As a first step, we looked into whether the individual accuracies of participants in the Stroop task and the Reasoning task were related. As Table 1 shows, although there was a slight trend toward a positive association between the final Stroop performance and the initial and final Reasoning performance, all correlations were weak and typically did not reach significance.

4.2.3.2. Direction of change

In order to obtain a more detailed picture of the relationship between the 2 tasks, we focused on the direction of change patterns (i.e., '00', '01', '10', '11'). This allowed us to examine whether the tendency to change one's response after deliberation (or not) was related in the 2 tasks. More specifically, for each direction of change category, we examined whether the proportion of trials of each category in the Stroop task was correlated with the proportion of trials of this same category in the Reasoning task. Table 2 shows the main results, but a full cross-tabulation table can also be found in Supplementary Material Section E.

As Table 2 shows, at the composite level, there was overall evidence for a weak positive association between the direction of change patterns of each task. The more a participant showed a specific change pattern in the Stroop task, the more they tended to show this pattern in the Reasoning task. At the composite level, the correlation reached significance for the '00' and '01' pattern. Hence, the more a reasoner tended to provide entirely incorrect responses in the Stroop task (i.e., both their initial and final responses were incorrect), the more they tended to do so in the Reasoning task. Likewise, the more a reasoner tended to correct an initial incorrect response after deliberation in the Stoop task, the more they tended to show this change pattern in the Reasoning task. For the '11' and '10' patterns the composite correlations did not reach significance. At the individual task level, the trends were more diffuse (Table 2).

To test whether the tendency to generate a correct final response through deliberation (i.e., a '01' pattern) showed a stronger link between the 2 tasks than the tendency to generate a correct response through intuitive processing (i.e., a '11' pattern) we also contrasted the composite '01' ($r = .18$) and '11' ($r = .12$) correlations directly. The difference between these correlations did not reach significance, $p = 0.61$.

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

Table 2. *Pearson's product–moment correlation between the proportion of each direction of change (i.e., '00', '01', '10', '00') of each individual in the Stroop task, and the proportion of each direction of change of that individual in the Reasoning task.*

	BB		BR		SYL		Composite	
	r	p	r	p	r	p	r	p
00	0.14	0.157	**0.19**	0.029	0.06	0.521	**0.17**	0.040
01	**0.20**	0.033	0.04	0.622	0.10	0.260	**0.17**	0.044
11	−0.02	0.817	0.04	0.654	0.14	0.092	0.12	0.149
10	−0.03	0.769	−0.05	0.598	**0.19**	0.022	0.12	0.161

Note: Correlations are reported both at the composite level and separately for each type of reasoning problem. 00, incorrect initial and incorrect final response; 01, incorrect initial and correct final response; 10, correct initial and incorrect final response; 11, correct final and correct initial response. Significant correlations ($p < .05$) are in bold. BB, bat-and-ball; BR, base-rates; SYL, syllogisms.

In sum, the correlational analyses indicated that there is evidence for a weak association between Stroop and Reasoning performance. However, there was no clear indication that initial Stroop performance would be a better predictor than the final Stroop performance or vice versa.

5. General discussion

In the present article, we were inspired by recent two-response findings that show evidence for correct intuitive responding in reasoning problems (e.g., Bago and De Neys, 2017, 2019a) and tested whether they could be generalized to low-level cognitive control tasks. For this purpose, we examined whether people who respond accurately to the classic Stroop and Flanker tasks could also do so when their deliberate control was minimized. We used the two-response paradigm to test the accuracy of both initial responses (given under limited deliberation conditions) and final responses. As a second step, we examined how the two-response Stroop performance was related to the performance on classic reasoning problems.

Concerning our first research question, both our studies showed that in most cases where people provided a correct final response to the Stroop and Flanker tasks, they had already responded correctly in the initial stage. In other words, deliberate control was not always necessary for correct responding in these tasks, which suggests that the two-response reasoning findings can generalize to lower-level cognitive control tasks. In general, this fits the claim that popular 'fast-and-slow' dual process models need to upgrade their view of the fast and intuitive System 1 (De Neys and Pennycook, 2019). Across a wide range of fields, responses that are traditionally believed to necessitate slow controlled deliberation, often seem to fall within the realm of more intuitive processing (De Neys, 2022).

As mentioned in the Introduction, the idea that control does not always require deliberation and can be exerted automatically, is in line with some existing evidence from the cognitive control field (Abrahamse et al., 2016; Chiu and Aron, 2014; Desender et al., 2013; Jiang et al., 2015, 2018; Linzarini et al., 2017). This evidence shows that participants perform better in an incongruent Stroop trial when it is preceded by an unconsciously presented incongruent trial (compared to an unconsciously presented congruent trial). In this case, participants recruit automatic control during the first, unconscious trial, which is boosting their performance in the trial that follows. These findings suggest that cognitive control, as we traditionally conceive it, might result from related automatic control processes, which fits with the findings of the present article. However, we would like to clarify that automatic ('intuitive') control is typically understood as unconscious control (e.g., implying subliminal presentation of trials). Although in the present article, our initial responses are extremely challenging for participants, they clearly fall outside the unconscious processing range. Therefore, we obviously do not argue that our

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

studies provide direct evidence for unconscious control, but that they point in the same direction as the aforementioned evidence of automatic control: control can be exerted faster and more effortlessly than traditionally assumed.

With our second research question, we attempted to explore how performance on cognitive control tasks, like the Stroop, and reasoning tasks are related. More specifically, we wanted to test whether an individual's initial Stroop performance would be a better predictor of their reasoning accuracy than their final Stroop performance. This question was inspired by Raoelison et al.'s (2020) research which showed that cognitive capacity primarily predicts intuitive, rather than deliberate reasoning performance. Under this 'smart intuitor' view, smarter people (i.e., people with high cognitive capacity) are better at providing correct responses intuitively, rather than deliberately correcting their erroneous intuitions. Our rationale was that both reasoning tasks and cognitive control tasks might tap into the same automatic control processes. This is why we expected that people who provide correct Stroop responses when their cognitive control is restricted, will also be able to provide correct intuitive responses to reasoning problems. However, we only found a weak association between people's response patterns in the Stroop task and those in the Reasoning tasks. Critically, there was no clear indication in our data to suggest that the initial, 'intuitive' Stroop performance could better predict reasoning accuracy compared to the final, deliberate Stroop performance. Below we discuss 2 main potential reasons for the lack of association between these tasks.

First, for practical reasons, in Study 3 of the present article, we focused on one cognitive control task, namely the Stroop. To our knowledge, our article is the first to test the corrective assumption in the Stroop task and investigate how it relates to reasoning accuracy. Thereby, it provides critical new insight into the generalization of the two-response findings. However, it is possible that the Stroop task alone was not an optimal psychometric predictor of cognitive control. While it is not uncommon to use a single task to tap cognitive control, a discussion exists in the literature concerning the task impurity problem (Miyake et al., 2000). More specifically, since no single cognitive control task is a pure measure of cognitive control, there are concerns that the observed results in studies that use only one type of predictor task are tied to the requirements of the task itself (e.g., specific demands and properties) rather than to cognitive control abilities (Gärtner and Strobel, 2021; Miyake et al., 2000). This might also explain why correlations of performance between these different cognitive control tasks are often weak or absent (e.g., Enge et al., 2014; Singh et al., 2018). Relatedly, deliberate control might not represent a single common process, but instead be separated into subtypes which would all be measured by different types of control tasks (e.g., Morra et al., 2018). For example, the Stroop task and the Flanker task that we used in Studies 1 and 2 are sometimes thought to tap into different aspects of cognitive control and measure different inhibition-related functions (e.g., Friedman and Miyake, 2004; Rey-Mermet et al., 2018; but also Nigg, 2000). While the Stroop task measures prepotent response inhibition (i.e., the ability to deliberately supress dominant or automatic responses), the Flanker task measures resistance to distractor interference (i.e., the ability to maintain focused attention and resist interference from distractors that are irrelevant to the task at hand). Although these 2 inhibitory functions are closely related (Friedman and Miyake, 2004), they are also distinguishable (Kane et al., 2016). One solution to combat these issues in future research would be to use a pool of common cognitive control tasks in order to create a cognitive control composite index. If we assume that our Stroop task is a weak indicator of individual cognitive control, this could explain why it is not strongly or differentially correlated with intuitive and deliberate reasoning performance.

Second, the weak association between the performance on the 2 tasks could also be due to their different nature. In the present article, we attempt to draw a link between the Stroop task and heuristics-and-biases reasoning tasks, but it might be that these are not necessarily directly comparable. That is, although the same pattern of results (i.e., correct intuitive responding) is present in both tasks, the specific mechanism that gives rise to this pattern might differ between them.

One of the potential differences between the Stroop task and heuristics-and-biases reasoning tasks is that in the reasoning tasks—for those who manage to respond correctly—the correct response might be more dominant because it is based on a rule that has been practiced to automaticity. That is, it has been

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

hypothesized that the origin of people's logical intuitions in reasoning tasks lies in a practice or learning process (De Neys, 2012; De Neys and Pennycook, 2019; Raoelison et al., 2021; Stanovich, 2018). Reasoners have typically already been exposed to the core logical principles and often even practiced them at length in the school curriculum (Raoelison et al., 2020). This repeated exposure would have allowed good reasoners to automatize their application. In other words, for sound reasoners the critical 'mindware' (i.e., the knowledge of elementary logical principles) has been fully instantiated (i.e., automatized, Stanovich, 2018) such that its activation strength will outcompete the conflicting heuristic intuition. In the Stroop task, however, the correct response is based on a new (in se trivial) instruction that people have not previously practiced or been exposed to. That is, participants are faced with an automatic, habitual response (i.e., reading the word) which they are told to consider as incorrect, and a competing response (i.e., naming the words' color) which they should consider as correct according to the task's instructions. Consequently, when responding to the Stroop task, participants always need to recruit cognitive control (automatically or deliberately) in order to inhibit their habitual response and answer correctly. However, because the more instantiated correct logical intuition will already dominate the competing heuristic intuition for good reasoners, correct intuitive responding may no longer require (or require less) control per se in a reasoning task. In sum, contrary to the Stroop task, correct responding to reasoning problems might not always demand engagement of (automatic) cognitive control, as the 2 potential responses are not always 'competing' with each other. This could explain why, in our findings, individual performance at the Stroop and the Reasoning task are not strongly related.

To sum up, we speculate that when solving reasoning problems one can be a sound intuitor either because one's 'logical' intuitions are very strong, or because one's competing logical and heuristic intuitions are similar in strength and cognitive control is automatically exerted (i.e., the heuristic response is automatically supressed). Interestingly, it has been argued in the reasoning field that the level of similarity between the alleged intuitions is reflected in response confidence: the more similar the competing intuitions are, the more conflicted and less certain one would feel about their decision (Bago and De Neys, 2020; De Neys, 2022; De Neys and Pennycook, 2019). This speculatively points to a possible test of this hypothesis. By contrasting initial correct responders who express the most and the least response confidence (i.e., who can be hypothesized to have less and more dominant logical intuitions, respectively), we can test whether the Reasoning-Stroop performance of the less confident (more conflicted) people is more strongly related. Presumably, participants with lower confidence have less dominant logical intuitions, thus they need automatic control to generate correct intuitive responses to reasoning problems (just like in the Stroop task). It is, therefore, in these participants that we may expect to find a clearer relationship between the initial Stroop and Reasoning performance.

Accidentally, we did (for different purposes, see Supplementary Material Section F) record response confidence for the Reasoning task in Study 3. We used these to split our group of reasoners into 2 halves based on the median '11' (i.e., correct final responses that were already generated intuitively) Reasoning task confidence: low '11' and high '11' confidence. We then performed a post hoc correlational analysis separately for each group. As it can be seen in Supplementary Material Section G, for the people that had a low '11' confidence (high conflict), their '11' Stroop performance clearly correlated with their '11' Reasoning composite performance (r = 0.34, p = .01). However, for the people that had a high '11' confidence (low conflict), their '11' Stroop performance did not correlate with their Reasoning performance (r = 0.07, p = .58). The difference between these correlations was not significant, p = .136. Although this post hoc analysis should be interpreted with caution, it does lend some credence to the idea that the Stroop and the Reasoning tasks might be only related in the cases where (automatic) cognitive control is required for sound reasoning.

Relatedly, it is worth considering that deliberation per se might play different roles in reasoning and lower-level cognitive control tasks. For example, in the reasoning field it has been shown that even when people intuitively arrive at the correct response, they subsequently engage in deliberation to justify that response (Bago and De Neys, 2019a; De Neys and Pennycook, 2019). In other words, although sound reasoners typically generate the correct response intuitively, they often struggle to explain how they arrived at their response (Bago and De Neys, 2019a). However, after the final response

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

stage, in which they are allowed to deliberate, they readily provide such justifications (e.g., Bago and De Neys, 2019a). Hence, it has been argued that deliberation during reasoning might be primarily required to justify and communicate one's response. Arguably, such justification is less central for lower-level cognitive control tasks. Although this hypothesis is speculative, it underscores that deliberation might play different or additional roles in these 2 domains.

A possible general critique against the present article is that we can never be sure that all possible deliberation was prevented in the initial, intuitive response stage. For example, it could be that the paradigm still allowed for some minimal deliberation during the initial stage, which could explain the correct responses at that stage. However, note that to minimize the possibility that reasoners engage in deliberate control in the initial stage, we combined 3 validated procedures which have been shown to reduce deliberation: instructions, time pressure, and concurrent load. One could always argue that a more demanding deadline or load task could have been used. Nevertheless, especially in the case of our low-level control tasks, it is important to consider the substantial number of missed trials both in Study 1 (41.9%), Study 2 (44.3%), and Study 3 (36.5%). These percentages suggest that the tasks were extremely challenging and that introducing additional load or time pressure would lead to practical and statistical issues (i.e., selection effects due to a large portion of discarded trials, e.g., Bouwmeester et al., 2017).

Nevertheless, the underlying point remains that regardless of how challenging the test conditions are in the initial stage, we can never be entirely certain that participants did not deliberate. The issue here is that dual process theories are underspecified (De Neys, 2021). While these theories suggest that deliberation is slower and more demanding than intuition, they do not provide a definite criterion or threshold for distinguishing between intuitive and deliberate processes (Bago and De Neys, 2019a; De Neys, 2022). So, as long as there are correct initial responses, one can always argue that they would disappear 'with just a little bit more load/time pressure'. At this point, the corrective assumption becomes unfalsifiable, since any evidence for correct intuiting can always be explained by arguing that the methodological design allowed for deliberation. At the same time, this indicates that the label correct 'intuiting' needs to be interpreted within practical boundaries and some caution. Although our results question the corrective role of deliberation in low-level control tasks, they should always be interpreted with this limitation in mind.

To conclude, although the link between cognitive control and reasoning performance might be complex, our key finding is that successful cognitive control does not necessarily require slow and effortful deliberation. This lends credence to the idea that cognitive control can be exerted automatically. These results point to an interesting generalization of the two-response findings to low-level cognitive control tasks. This further underscores the claim that the popular 'fast-and-slow' dual process models of human cognition need to revise and upgrade their view of the fast and intuitive System 1 (De Neys and Pennycook, 2019). We also hope that the study can serve as a proof-of-principle and lead to a deeper integration of the related—but hitherto somewhat isolated—cognitive control and reasoning fields. We believe that such an integration will be indispensable to pinpoint the mechanisms underlying intuitive-automatic responding in higher- and lower-level cognition.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/jdm.2023.32.

Funding statement. This research was supported by a research grant (DIAGNOR, ANR16-CE28-0010-01) from the Agence Nationale de la Recherche, France. A.V. was supported by an IdEx doctoral fellowship from the Université Paris Cité.

Competing interest. No potential competing interest was reported by the authors.

References

Abrahamse, E., Braem, S., Notebaert, W., & Verguts, T. (2016). Grounding cognitive control in associative learning. *Psychological Bulletin*, *142*(7), 693–728. https://doi.org/10.1037/bul0000047

Abreu-Mendoza, R. A., Coulanges, L., Ali, K., Powell, A. B., & Rosenberg-Lee, M. (2020). Children's discrete proportional reasoning is related to inhibitory control and enhanced by priming continuous representations. *Journal of Experimental Child Psychology*, *199*, 104931. https://doi.org/10.1016/j.jecp.2020.104931

Abutalebi, J., Della Rosa, P. A., Green, D. W., Hernandez, M., Scifo, P., Keim, R., . . . Costa, A. (2012). Bilingualism tunes the anterior cingulate cortex for conflict monitoring. *Cerebral Cortex*, *22*(9), 2076–2086. https://doi.org/10.1093/cercor/bhr287

Aïte, A., Cassotti, M., Linzarini, A., Osmont, A., Houdé, O., & Borst, G. (2016). Adolescents' inhibitory control: Keep it cool or lose control. *Developmental Science*, *21*(1), e12491. https://doi.org/10.1111/desc.12491

Algom, D., & Chajut, E. (2019). Reclaiming the stroop effect back from control to input-driven attention and perception. *Frontiers in Psychology*, *10*, 1683. https://doi.org/10.3389/fpsyg.2019.01683

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. 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. 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

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624

Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., . . . Wollbrant, C. E. (2017). Registered replication report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, *12*(3), 527–542. https://doi.org/10.1177/1745691617693624

Braem, S., & Egner, T. (2018). Getting a grip on cognitive flexibility. *Current Directions in Psychological Science*, *27*(6), 470–476. https://doi.org/10.1177/0963721418787475

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

Chiu, Y.-C., & Aron, A. R. (2014). Unconsciously triggered response inhibition requires an executive setting. *Journal of Experimental Psychology*, *143*(1), 56–61. https://doi.org/10.1037/a0031497

de Fockert, J. W., Rees, G., Frith, C. D., & Lavie, N. (2001). The role of working memory in visual selective attention. *Science*, *291*(5509), 1803–1806. https://doi.org/10.1126/science.1056496

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, *17*(5), 428–433. https://doi.org/10.1111/j.1467-9280.2006.01723.x

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. (2017). Bias, conflict, and fast logic: Towards a hybrid dual process future? In *Dual Process Theory 2.0*. Oxon: 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. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, *46*, E111. https://doi.org/10.1017/S0140525X2200142X

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS One*, *6*(1), e15954. https://doi.org/10.1371/journal.pone.0015954

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

Desender, K., Lierde, E. V., & Bussche, E. V. (2013). Comparing conscious and unconscious conflict adaptation. *PLoS One*, *8*(2), e55976. https://doi.org/10.1371/journal.pone.0055976

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Enge, S., Behnke, A., Fleischhauer, M., Küttler, L., Kliegel, M., & Strobel, A. (2014). No evidence for true training and transfer effects after inhibitory control training in young healthy adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 987. https://doi.org/10.1037/a0036165

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Eriksen, C. W., & Hoffman, J. E. (1973). The extent of processing of noise elements during selective encoding from visual displays. *Perception & Psychophysics*, *14*(1), 155–160. https://doi.org/10.3758/BF03198630

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

Evans, J. St. B. T. (2002). Logic and human reasoning: An assessment of the deduction paradigm. *Psychological Bulletin*, *128*(6), 978–996. https://doi.org/10.1037/0033-2909.128.6.978

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, Judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans J. St. 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. https://doi.org/10.1080/13546783.2019.1623071

Evans, J. St. B. T., & Over, D. E. (1996). Rationality in the selection task: Epistemic utility versus uncertainty reduction. *Psychological Review*, *103*(2), 356–363. https://doi.org/10.1037/0033-295X.103.2.356

Fan, J., McCandliss, B. D., Fossella, J., Flombaum, J. I., & Posner, M. I. (2005). The activation of attentional networks. *Neuroimage*, *26*(2), 471–479. https://doi.org/10.1016/j.neuroimage.2005.02.004

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning*, *15*(2), 105–128. https://doi.org/10.1080/13546780802711185

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133*(1), 101. https://doi.org/10.1037/0096-3445.133.1.101

Gao, Q., Chen, Z., & Russell, P. (2007). Working memory load and the stroop interference effect. *New Zealand Journal of Psychology 36*(3), 146–153. http://hdl.handle.net/10092/2792

Gärtner, A., & Strobel, A. (2021). Individual differences in inhibitory control: A latent variable analysis. *Journal of Cognition*, *4*(1), 17. https://doi.org/10.5334/joc.150

Handley, S. J., Capon, A., Beveridge, M., Dennis, I., & Evans, J. S. B. (2004). Working memory, inhibitory control and the development of children's reasoning. *Thinking & Reasoning*, *10*(2), 175–195. https://doi.org/10.1080/13546780442000051

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. https://doi.org/10.1037/a0021098

Hübner, R., Steinhauser, M., & Lehle, C. (2010). A dual-stage two-phase model of selective attention. *Psychological Review*, *117*(3), 759. https://doi.org/10.1037/a0019471

Jiang, J., Correa, C. M., Geerts, J., & van Gaal, S. (2018). The relationship between conflict awareness and behavioral and oscillatory signatures of immediate and delayed cognitive control. *NeuroImage*, *177*, 11–19. https://doi.org/10.1016/j.neuroimage.2018.05.007

Jiang, J., Zhang, Q., & Van Gaal, S. (2015). EEG neural oscillatory dynamics reveal semantic and response conflict at difference levels of conflict awareness. *Scientific Reports*, *5*(1), 12008. https://doi.org/10.1038/srep12008

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, *164*, 56–64. https://doi.org/10.1016/j.actpsy.2015.12.008

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

Kahneman, D., & Chajczyk, D. (1983). Tests of the automaticity of reading: Dilution of Stroop effects by color-irrelevant stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *9*(4), 497. https://doi.org/10.1037/0096-1523.9.4.497

Kane, M. J., Meier, M. E., Smeekens, B. A., Gross, G. M., Chun, C. A., Silvia, P. J., & Kwapil, T. R. (2016). Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General*, *145*(8), 1017. https://doi.org/10.1037/xge0000184

Keele, S. W. (1972). Attention demands of memory retrieval. *Journal of Experimental Psychology*, *93*(2), 245. https://doi.org/10.1037/h0032460

Kessler, J., Kivimaki, H., & Niederle, M. (2017). Thinking fast and slow: Generosity over time. Preprint at https://stanford.edu/~niederle/KKN_ThinkingFastandSlow.pdf

Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in Cognitive Sciences*, *9*(2), 75–82. https://doi.org/10.1016/j.tics.2004.12.004

Lavie, N., & De Fockert, J. (2005). The role of working memory in attentional capture. *Psychonomic Bulletin & Review*, *12*(4), 669–674. https://doi.org/10.3758/BF03196756

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, *133*(3), 339–354. https://doi.org/10.1037/0096-3445.133.3.339

Linzarini, A., Houdé, O., & Borst, G. (2017). Cognitive control outside of conscious awareness. *Consciousness and Cognition*, *53*, 185–193. https://doi.org/10.1016/j.concog.2017.06.014

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. (2018). Response: Commentary: Seeing the conflict: an attentional account of reasoning errors. *Frontiers in Psychology*, *9*, 24. https://doi.org/10.3389/fpsyg.2018.00024

Mata, A., Schubert, A. L., & Ferreira, M. B. (2014). The role of language comprehension in reasoning: How "good-enough" representations induce biases. *Cognition*, *133*(2), 457–463. https://doi.org/10.1016/j.cognition.2014.07.011

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

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

https://doi.org/10.1017/jdm.2023.32 Published online by Cambridge University Press

Morra, S., Panesi, S., Traverso, L., & Usai, M. C. (2018). Which tasks measure what? Reflections on executive function development and a commentary on Podjarny, Kamawar, and Andrews (2017). *Journal of Experimental Child Psychology*, *167*, 246–258. https://doi.org/10.1016/j.jecp.2017.11.004

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. https://doi.org/10.1037/xlm0000372

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin*, *126*(2), 220. https://doi.org/10.1037/0033-2909.126.2.220

Penner, I. K., Kobel, M., Stöcklin, M., Weber, P., Opwis, K., & Calabrese, P. (2012). The Stroop task: comparison between the original paradigm and computerized versions in children and adults. *The Clinical Neuropsychologist*, *26*(7), 1142–1153. https://doi.org/10.1080/13854046.2012.713513

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. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001

Raoelison, M., Boissin, E., Borst, G., & De Neys, W. (2021). From slow to fast logic: the development of logical intuitions. *Thinking & Reasoning*, *27*(4), 599–622. https://doi.org/10.1080/13546783.2021.1885488

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.

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

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 501. https://doi.org/10.1037/xlm0000450

Ridderinkhof, K. R., Wylie, S. A., van den Wildenberg, W. P., Bashore, T. R., & van der Molen, M. W. (2021). The arrow of time: Advancing insights into action control from the arrow version of the Eriksen flanker task. *Attention, Perception, & Psychophysics*, *83*, 700–721. https://doi.org/10.3758/s13414-020-02167-z

Servant, M., & Logan, G. D. (2019). Dynamics of attentional focusing in the Eriksen flanker task. *Attention, Perception, & Psychophysics*, *81*, 2710–2721. https://doi.org/10.3758/s13414-019-01796-3

Singh, K., Ecker, U., Gignac, G., Brydges, C., & Rey-Mermet, A. (2018). *Interference control in working memory*. PsyArXiv. https://doi.org/10.31234/osf.io/fjrnq

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. *Behavioral and Brain Sciences*, *23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Stirling, N. (1979). Stroop interference: An input and an output phenomenon. *The Quarterly Journal of Experimental Psychology*, *31*(1), 121–132. https://doi.org/10.1080/14640747908400712

Stoffels, E. J., & Van der Molen, M. W. (1988). Effects of visual and auditory noise on visual choice reaction time in a continuous-flow paradigm. *Perception & Psychophysics*, *44*, 7–14. https://doi.org/10.3758/bf03207468

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test-retest reliability of standard and emotional stroop tasks: An investigation of color-word and picture-word versions. *Assessment*, *12*(3), 330–337. https://doi.org/10.1177/1073191105276375

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643. https://doi.org/10.1037/h0054651

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

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

Wright, B. C., & Wanley, A. (2003). Adults' versus children's performance on the Stroop task: Interference and facilitation. *British Journal of Psychology*, *94*(4), 475–485. https://doi.org/10.1348/000712603322503042

Cite this article: Voudouri, A., Bago, B., Borst, G., and De Neys, W. (2023). Reasoning and cognitive control, fast and slow. *Judgment and Decision Making*, e33. https://doi.org/10.1017/jdm.2023.32

Supplementary Material

A. Instructions

Stroop task instructions

The literal instructions that were used in the two-response Stroop task stated the following:

> “Welcome to the experiment! This experiment will take about 30 minutes to complete and it demands your full attention. You can only do this experiment once. Click on Next to start. Please read these instructions carefully! In this task you will be presented with words, one after the other, to the centre of the screen, and you need to respond to the colour that each word is presented in. Press: d for red; f for blue; j for green; k for yellow. You can see an example of the words below. In this example you would have to press f for blue.”

The word “blue” written in blue ink colour was displayed on screen.

> “We are going to start with a couple of practice problems to familiarise you with the buttons.
> In this practice you will only be presented with colours, not words. First, a fixation cross will appear. Then a colour will appear and you will need to click on the corresponding button. Please respond as fast and as accurately as possible (try to answer as fast as you can while not making mistakes). After you respond, you will be given feedback for your responses. Once you click on the button, you will be automatically taken to the next page. Remember: Press d for red; f for blue; j for green; k for yellow. Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. Press SPACE to start the practice”
>
> “This is the end of this practice. Now you are going to practice with the words. You need to respond to the colour that each word is presented in. Please respond as fast and as accurately as possible (try to answer as fast as you can while not making mistakes). You will be given feedback for your responses. Remember: Press d for red; f for blue; j for green; k for yellow
> Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. In the actual experiment, sometimes the ink color in which the word appears will not match with the word. For example, the following word could appear:”

The word “green” written in yellow ink was displayed on screen.

> “Here the word "green" is written in yellow. We ask you to always respond to the color of the word. So in this example you would need to press the button ‘k’ for ‘yellow’. We will let you practice a couple of these now. You will get feedback for your responses. Press d for red; f for blue; j for green; k for yellow. Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. Press SPACE when you are ready to start the practice.”
>
> “This is the end of this practice. In the actual task, you will give two responses to each word. First, we want to know what your initial, intuitive response to the colour of each word is and afterwards we want to see how you respond after you have thought about the colour of each word for some more time. So, for the first response you need to give 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 make sure that you answer as fast as possible, a time limit is set for the first response, which is going to be 750 milliseconds (that's less than a second!). Please make sure to answer before the deadline passes. In the next part, you are going to watch an initial trial to get a feel of the deadline. Press Next to see the trial.”

> “This is how fast the word is going to be presented! You need to give a response within this time. You are now going to practice this with some words. First, a fixation cross will appear. Then the word will appear and you will need to click on the button that corresponds to the colour of the word. As we mentioned before, we are first interested in your initial, intuitive response. Next, the word will be presented again and you can take all the time you want to actively reflect on your choice. Once you have made up your mind you enter your final response. After you click on the button, you will be automatically taken to the next page. From here on we will no longer tell you whether the color you picked was correct or not. We will let you know whenever you responded too slowly and missed the deadline. Remember: Press d for red; f for blue; j for green; k for yellow. Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. Press SPACE to start the practice.”

> “This is the end of this practice. In the actual task, you will also need to memorise six numbers while you respond to the words. The numbers will be displayed for 2 seconds and then you will view one number with a question mark. You have to press 'd' for yes, the number was part of the set, or press 'k' for no, the number was not part of the set. There is no deadline for your response. You will get feedback after each response. To better understand this, you will first practise with five sets of numbers without the words. You should prepare yourself by holding the index finger of your left hand on the "d" key and the index finger of your right hand over the "k" key. Press SPACE to begin.”

> “In the actual task you will need to memorise the numbers while you respond to the words. The numbers will be briefly presented before each word. We know that it is not always easy to memorise the numbers while you are also thinking about the words. The most important thing is to correctly memorise the numbers. First, try to concentrate on the memorisation task, and then try to solve the colour-word task. The memorization will only be required for your first, intuitive response. For your final response you can take as much time as you want without having to memorize the pattern. You can practice this in this practice round. Remember: Press d for red; f for blue; j for green; k for yellow. Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. Press SPACE to continue to the practice”

> “This is the end of all practice rounds! Now you will begin with the task. In the colour-word task there will be a total of 128 trials grouped in 3 blocks. After each block you can take a short break. Within each block one trial will be presented immediately after the other and you should not pause between them. In total the 3 blocks will take approximately 15 minutes. Please make sure to stay maximally focused throughout the study. Remember: Press d for red; f for blue; j for green; k for yellow. Prepare yourself by holding the middle and index fingers of your left hand on the "d" and "f" keys and the middle and index fingers of your right hand over the "j" and "k" keys, like it is shown below. Press SPACE when you’re ready to start with the first block”

> “BREAK You just finished the first block! There are two blocks remaining. Feel free to take a short break. Before you start remember: Press d for red; f for blue; j for green; k for yellow. Prepare yourself by placing the middle and index fingers of your left hand on the ‘d’ and ‘f’ keys and the middle and index fingers of your right hand over the ‘j’ and "k" keys, like it is shown below. Press SPACE when you are ready to continue to the next block.”

Flanker task instructions

The literal instructions that were used in the two-response Flanker task stated the following:

> “Welcome to the experiment!
> This experiment will take about 24 minutes to complete and it demands your full attention. You can only do this experiment once. Click on Next to start. Please read these instructions carefully! In this task you will be presented with an arrow at the center of the screen, which will look like the arrows that are shown below.”

Two arrows, one pointing to the left and one to the right, were displayed on the screen.

> “Your task will be to press the button that matches the direction the arrow is pointing to. Click on Next to continue. Press F if the central arrow is pointing Left. Press the correct key to continue. Press J if the arrow is pointing Right. Press the correct key to continue”

Two rows of five arrows were displayed on the screen, one after the other.

> “The central arrow will always be presented along with four other arrows as it is shown below. Your task is to identify the direction of the CENTRAL arrow. Ignore the peripheral arrows. Remember: Press F if the central arrow is pointing Left. Press J if the central arrow is pointing Right. Click on Next to continue.”

> “We are going to start with 6 practice trials to familiarise you with the buttons. First, a fixation cross will appear. Then five arrows will appear and you should identify the direction of the CENTRAL arrow by clicking on the corresponding button. Remember: Press F if the central arrow is pointing Left. Press J if the central arrow is pointing Right. You should prepare yourself by holding the index finger of your left hand on the F key and the index finger of your right hand on the J key. After you respond, you will be given feedback for your responses. Once you click on a key, you will be automatically taken to the next trial. Press SPACE to start the practice.”

> “This is the end of this practice. In the actual task, you will give two responses to each trial. First, we want to know what your initial, intuitive response to the direction of the central arrow is and afterwards we want to see how you respond after you have thought about it for some more time. So, for the first response you need to give 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 make sure that you answer as fast as possible, a time limit is set for the first response, which is going to be 420 milliseconds (that's less than half a second!). Please make sure to answer before the deadline passes. In the next part, you are going to watch an initial trial to get a feel of the deadline. Press Next to see the trial.”

After a fixation cross was shown, a row of five arrows was displayed on the screen.

> “This is how fast the word is going to be presented! You need to give a response within this time. You are now going to practice this with some trials. First, a fixation cross will appear. Then the arrows will appear and you will need to click on the button that corresponds to the direction of the central arrow. As we mentioned before, we are first interested in your initial, intuitive response. Next, you will see the reminder “Please give your final response”. The same arrows will be presented again and you can take all the time you want to actively reflect on the direction of the central arrow. Once you have made up your mind you can enter your final response. After you click on the key, you will be automatically taken to the next trial. We will no longer tell you whether the direction you picked was correct or not. We will only let you know whenever you responded too slowly and missed the deadline. Remember: Press F if the central arrow is pointing Left. Press J if the central arrow is pointing Right. You should prepare yourself by holding the index finger of your left hand on the F key and the index finger of your right hand on the J key. Press SPACE to start this practice session.”

> “This is the end of this practice. In the actual task, you will also need to memorise six numbers while you view the arrows. The numbers will be displayed for 2 seconds and then you will view one number with a question mark. You have to press F for yes, the number was part of the set, or press J for no, the number was not part of the set. There is no deadline for your response. You will get feedback after each response. To better understand this, you will first practise with five sets of numbers without the arrows. You should prepare yourself by holding the index finger of your left hand on the F key and the index finger of your right hand over the J key. Press SPACE to begin.”

> “In the actual task you will need to memorise the numbers while you respond to the direction of the central arrow. The numbers will be briefly presented before the arrows. We know that it is not always easy to memorise the numbers while you are also thinking about the direction of the central arrow. The most important thing is to correctly memorise the numbers. First, try to concentrate on the memorisation task, and then try to solve the arrow task. The memorization will only be required for your first, intuitive response. For your final response you can take as much time as you want without having to memorize the pattern. You can practice this in this practice round.”

> “This is the end of all practice rounds! Now you will begin with the task. There will be a total of 128 trials grouped in 3 blocks. After each block you can take a short break. Within each block one trial will be presented immediately after the other and you should not pause between them. In total the 3 blocks will take approximately 18 minutes. Please make sure to stay maximally focused throughout the study. Remember: Press F

if the central arrow is pointing Left. Press J if the central arrow is pointing Right. You should prepare yourself by holding the index finger of your left hand on the F key and the index finger of your right hand over the J key. Press SPACE when you’re ready to start with the first block.”

“BREAK You just finished the first block! There are two blocks remaining. Feel free to take a short break. Before you start remember: Press F if the central arrow is pointing Left. Press J if the central arrow is pointing Right. You should prepare yourself by holding the index finger of your left hand on the F key and the index finger of your right hand over the J key. Press SPACE when you are ready to continue to the next block.”

B. Reaction times

Table S1.

Mean (SD) of reaction times in Study 1 (Stroop task), Study 2 (Flanker task) and Study 3 (Stroop task) as a function of congruency status (congruent; incongruent), Response stage (initial response; final response) and response accuracy (correct; incorrect; overall). Reaction times are expressed in milliseconds. The first column ("Overall") refers to both correct and incorrect trials combined.

		Overall		Correct		Incorrect	
		Incongruent	Congruent	Incongruent	Congruent	Incongruent	Congruent
Study 1	Initial	581 (55)	542 (104)	577 (56)	557 (65)	586 (66)	532 (133)
	Final	982 (744)	890 (873)	1019 (890)	895 (884)	942 (786)	769 (463)
Study 2	Initial	315 (44)	306 (58)	316 (48)	315 (56)	301 (43)	253 (66)
	Final	543 (260)	515 (225)	529 (210)	516 (225)	484 (331)	420 (181)
Study 3	Initial	580 (53)	545 (53)	576 (54)	547 (48)	592 (72)	542 (89)
	Final	1096 (2536)	764 (640)	1078 (2524)	765 (669)	1166 (2709)	831 (523)

C. Inclusion of all trials

Table S2.

Direction of change proportions (%) by Congruency status (congruent; incongruent) in Study 1 (Stroop task), Study 2 (Flanker task) and Study 3 (Stroop task) including all missed load and missed deadline trials. All missed deadline trials were coded as “0” (i.e., incorrect response).

		“00”	“01”	“10”	“11”
Study 1	Incongruent	10.5	47.2	1.1	41.2
	Congruent	2.8	35.3	1.2	60.7
Study 2	Incongruent	7.2	52.3	1.3	39.2
	Congruent	1.2	41.3	0.4	57.2
Study 3	Incongruent	8.7	46.2	1.9	43.2
	Congruent	2.5	32.2	1.7	63.7

Note. “00” = incorrect initial and incorrect final response; “01” = incorrect initial and correct final response; “10” = correct initial and incorrect final response; “11” = correct final and correct initial response.

D. Stroop task results of Study 3

Accuracy

As Figure S1A shows, we replicated the key pattern of results that we observed in Study 1. When participants were allowed to deliberate, they typically managed to solve incongruent trials correctly, but they still performed better on congruent compared to incongruent trials. The mean accuracy for the initial responses of the congruent trials was 79.8% (SD = 18.2%) and differed from 25% chance (t(140) = 35.87, p < .001). The mean accuracy for the initial responses of the critical incongruent trials was 63.6% (SD = 24.3%) and also differed from 25% chance (t(138) = 18.67, p < .001). This suggests that even when participants were forced to rely on intuitive, automatic processing, they were often able to produce correct responses. To see if there was an effect of the response stage (initial; final) and the congruency status (congruent; incongruent) on the accuracy of the Stroop responses, a two-way within-subjects ANOVA was conducted. As Figure S1A shows, the accuracy for congruent trials was higher than for incongruent trials (F(1, 137) = 70.41, p < .001, η^2g = 0.103) and the accuracy at the final stage was higher than at the initial stage (F(1, 137) = 275.40, p < .001, η^2g = 0.287), indicating that accuracy improved after deliberation. Finally, the difference between initial and final accuracy was higher for incongruent compared to congruent trials, as indicated by the response stage by congruency interaction (F(1, 137) = 60.93, p < .001, η^2g = 0.287).

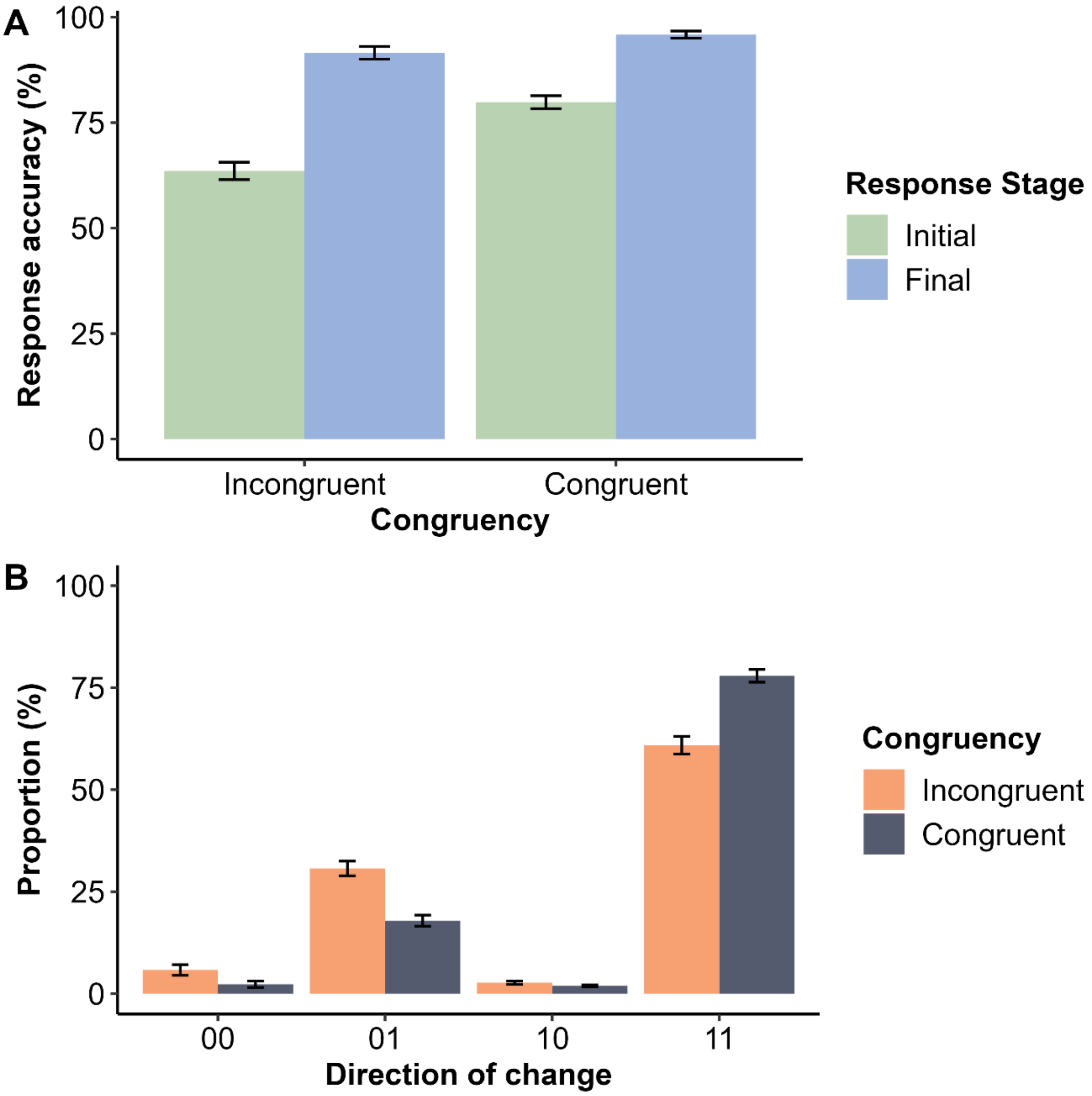

Figure S1. Accuracy and Direction of Change in the Stroop task of Study 3. **A)** Response accuracy at incongruent and congruent trials as a function of response stage. **B)** Proportion of each direction of change category in incongruent and congruent trials. The error bars represent the Standard Error of the Mean. “00” = incorrect initial and incorrect final response; “01” = incorrect initial and correct final response; “10” = correct initial and incorrect final response; “11” = correct final and correct initial response.

Stability index

The average stability index for the initial responses of the critical, incongruent trials was 74.2% (SD = 13.8%). If initial responding was prone to systematic guessing, we would expect more inconsistency in participants’ initial responses across trials.

Direction of change

To get a more precise picture of how participants changed their responses after deliberation we conducted a direction of change analysis (Bago & De Neys, 2017, 2019a). The proportions of each direction of change were very similar to those of Study 1. As Figure S1B shows, the vast majority of the critical, incongruent trials had a “11” pattern (60.9%). The high “11” proportion was accompanied by a low “00” proportion (5.8%), and “10” proportion (2.7%). More importantly, the proportion of “11” trials was higher than that of the “01” trials (30.7%). The non-correction rate (i.e., proportion 11/11+01) reached 66.5%. This confirms the results of Study 1 and indicates that, in most correct final trials, the correct response was already generated when deliberate control was minimized.

As Figure S1B shows, and as it was expected, a similar pattern was observed for congruent trials. In most trials, correct responses were intuitively generated and the non-correction rate reached 81.3%.

Reaction Times

The average reaction time at the initial response stage was 545 ms (SD = 53 ms) for the congruent trials, and 580 ms (SD = 53 ms) for the incongruent trials. Participants spent longer on the final response stage, with an average of 764 ms (SD = 640 ms) at congruent trials and 1096 ms (SD = 2536 ms) at incongruent trials. Supplementary Material section B gives a full overview of reaction times according to response accuracy.

Exploratory analysis

To make maximally sure that participants did not deliberate during the initial response stage, we excluded a considerable amount of trials. As mentioned in Study 1, this could have artificially boosted the critical non-correction rate. To examine this possibility, we re-ran the direction of change analysis while including all missed load and missed deadline trials. As in Study 1, we opted for the strongest possible test and coded the accuracy of all missed deadline trials as “0” (i.e., incorrect). In the missed load trials both initial and final responses were recorded. The analysis, as reported in Supplementary Material section C, pointed to a higher proportion of “01” incongruent trials (46.2%), but the proportion of “11” (43.2%) responses and the non-correction rate remained high (48.3%). As in Study 1, even in this extremely conservative analysis, correct incongruent responses were still generated intuitively about half of the time.

To summarize, regarding the Stroop task, the results of Study 3 replicated those of Study 1, with a much larger sample. This confirms the main finding of Study 1: even when deliberate control is minimized, participants can typically still provide correct Stroop responses. This suggest that more often than not, correct responding on the Stroop trial seems to be done intuitively in the absence of deliberate controlled correction.

E. Full cross tabulation table of correlation

Table S3

Pearson's product-moment correlation tests between the proportion of each direction of change (i.e., “00”, “01”, “10”, “00”) of each individual at the Stroop task, and the proportion of each direction of change of that individual at the Reasoning task of Study 3. Correlations are reported both at the composite level and separately for each type of reasoning problem.

Direction Reasoning	Task	Direction Stroop							
		00		01		10		11	
		r	p	r	p	r	p	r	p
00	BB	0.14	0.157	**−0.23**	0.016	0.11	0.251	0.07	0.439
	BR	**0.19**	0.029	−0.05	0.581	0.13	0.137	−0.10	0.252
	SYL	0.06	0.521	−0.04	0.676	0.03	0.743	−0.01	0.902
	Composite	**0.17**	0.040	−0.09	0.313	0.12	0.166	−0.06	0.503
01	BB	−0.08	0.432	**0.20**	0.033	−0.06	0.545	−0.11	0.276
	BR	−0.08	0.333	0.04	0.622	−0.11	0.206	0.04	0.675
	SYL	0.12	0.169	0.10	0.260	−0.07	0.404	−0.14	0.108
	Composite	−0.55	0.949	**0.17**	0.044	−0.12	0.147	−0.11	0.180
10	BB	−0.06	0.516	0.09	0.326	−0.03	0.769	−0.03	0.760
	BR	−0.07	0.433	−0.13	0.142	−0.05	0.598	0.15	0.073
	SYL	0.07	0.439	0.14	0.098	**0.19**	0.022	**−0.19**	0.023
	Composite	−0.02	0.781	0.07	0.427	0.12	0.161	−0.06	0.457
11	BB	−0.12	0.227	0.15	0.122	−0.10	0.291	−0.02	0.817
	BR	−0.5	0.136	0.06	0.475	−0.06	0.472	0.04	0.654
	SYL	−0.13	0.128	−0.06	0.464	−0.06	0.449	0.14	0.092
	Composite	**−0.18**	0.038	0.76	0.930	−0.12	0.172	0.12	0.149

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; “01” = incorrect initial and correct final response; “00” = incorrect initial and incorrect final response; “10” = correct initial and incorrect final response; “11” = correct final and correct initial response. Significant correlations (*p* < .05) are in bold.

F. Reasoning confidence

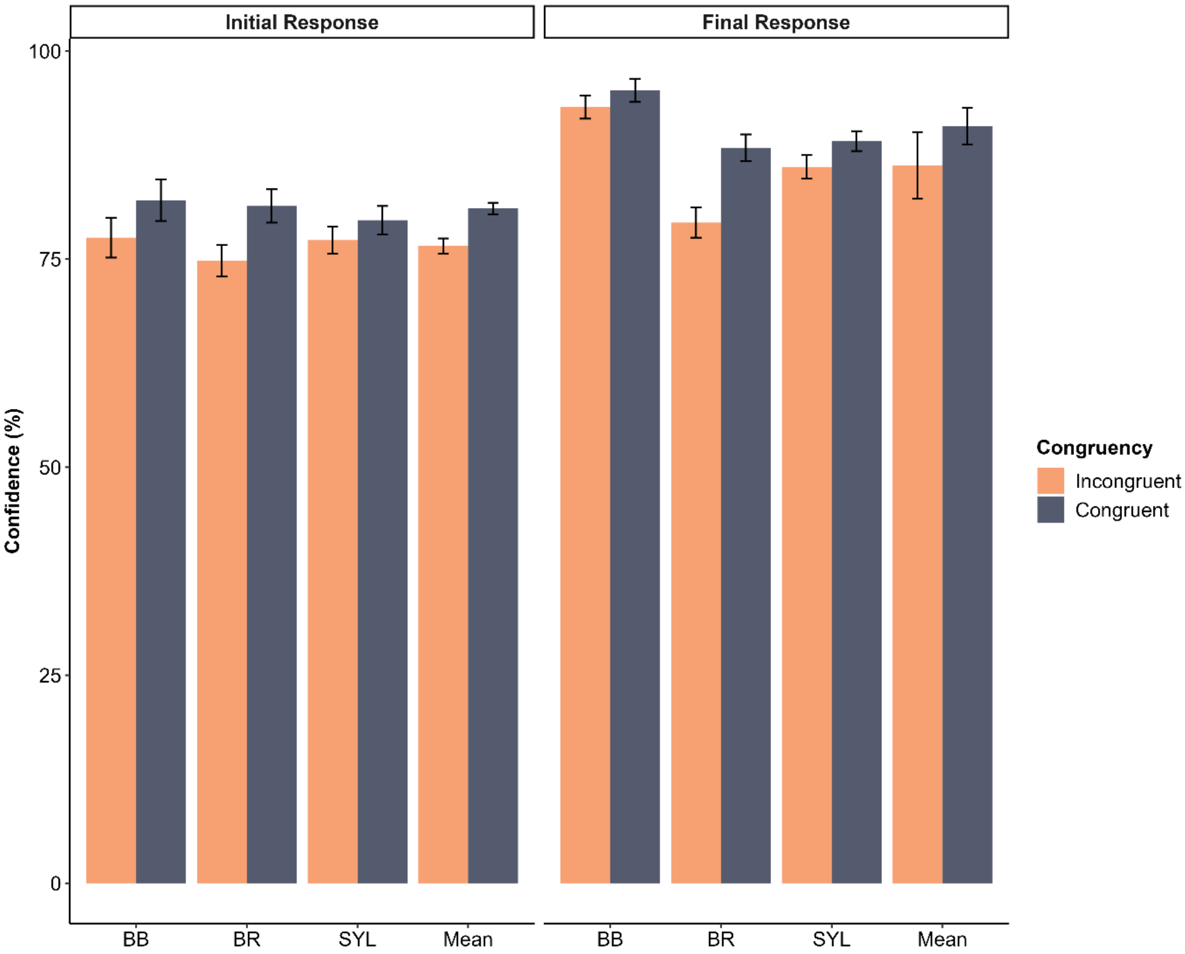

Figure S2. Confidence at initial and final responses, at congruent and incongruent trials in the Reasoning task of Study 3, separately for each problem type and for the mean across the three problem types. The error bars represent the Standard Error of the Mean. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; Mean = the mean across the four tasks.

Table S4.

Mean (SD) of the reported confidence at the initial responses of the Reasoning task of Study 3, as a function of congruency status (Congruent; Incongruent) and problem type (BB; BR; SYL; Mean).

	BB	BR	SYL	Mean
Incongruent	77.6 (28.9)	74.8 (26.4)	77.3 (25.4)	76.6 (1.5)
Congruent	82.1 (30.2)	81.4 (26.3)	79.7 (26.0)	81.1 (1.2)

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; Mean = the mean across tasks.

G. Correlation results according to “11” conflict level

Conflict detection in the Reasoning task of Study 3 was calculated by subtracting the baseline confidence (i.e., the confidence at the correct congruent trials), from the confidence at the incongruent trials (e.g., De Neys et al., 2013; 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 reasoner in the incongruent trials. In sum, high conflict detection is equivalent to low response confidence.

Table S5

Summary statistics of the initial conflict detection at the “11” trials of the Reasoning task (Study 3), separately for the half of the group that had a high conflict detection at “11” trials (“High half”) and for the half of the group that had a low conflict detection at “11” trials (“Low half”). Negative values point to an overall successful conflict detection.

	N	Min	Max	Median	Q1	Q3	IQR	Mad	Mean	SD	SE	CI
High half	58	−100	-5	−16.86	−24.69	−12.5	12.19	9.33	−21.17	15.64	2.05	4.11
Low half	58	−3.87	33.33	0.44	0	9.67	9.67	3.12	5.31	8.84	1.16	2.32

Table S6

Correlation tests between the proportion of each direction of change of each individual at the Stroop task (Study 3), and the proportion of each direction of change of that individual at the Reasoning task (Study 3), for the half of the participants that had a high conflict detection at “11” trials.

	BB		BR		SYL		Composite	
	r	*p*	r	*p*	r	*p*	r	*p*
00	0.19	0.226	0.25	0.061	−0.30	0.841	0.22	0.109
01	**0.34**	0.029	***0.41***	0.002	0.003	0.982	**0.35**	0.009
11	0.10	0.524	**0.29**	0.029	0.23	0.088	**0.34**	0.010
10	−0.06	0.718	−0.07	0.617	**0.33**	0.014	0.18	0.185

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; “00” = incorrect initial and incorrect final response; “01” = incorrect initial and correct final response; “10” = correct initial and incorrect final response; “11” = correct final and correct initial response. Significant correlations at the 0.05 level are in bold. Significant correlations at the 0.01 level are in bold and italics.

Table S7

Correlation tests between the proportion of each direction of change of each individual at the Stroop task (Study 3), and the proportion of each direction of change of that individual at the Reasoning task (Study 3), for the half of the participants that had a low conflict detection at “11” trials.

	BB		BR		SYL		Composite	
	r	*p*	r	*p*	r	p	r	*p*
00	0.15	0.332	0.25	0.053	0.16	0.230	0.25	0.056
01	0.12	0.450	−0.17	0.211	**0.34**	0.009	0.15	0.248
11	−0.07	0.664	−0.13	0.349	0.25	0.060	0.07	0.581
10	0.05	0.985	−0.12	0.357	0.07	0.604	0.002	0.985

Note. BB = Bat-and-ball; BR = Base-rates; SYL = Syllogisms; “00” = incorrect initial and incorrect final response; “01” = incorrect initial and correct final response; “10” = correct initial and incorrect final response; “11” = correct final and correct initial response. Significant correlations (*p* < .05) are in bold.