Thinking & Reasoning, in press

From slow to fast logic: The development of logical intuitions

Matthieu Raoelison[1], Esther Boissin[1],
GrégoireBorst[1], Wim De Neys[1]

[1]Université de Paris, LaPsyDÉ, CNRS, F-75005 Paris, France

Abstract

Recent reasoning accounts suggest that people can process elementary logical principles intuitively. These controversial "logical intuitions" are believed to result from a learning process in which developing reasoners automatize their application. To verify this automatization hypothesis, we contrasted the reasoning performance of younger (7th grade) and older (12th grade) reasoners with a two-response paradigm. Participants initially responded with the first intuitive response that came to mind and subsequently were allowed to deliberate on classic "bias" problems (base-rate problems and syllogisms). Results showed that in addition to showing less deliberate correction of an initial erroneous response, younger reasoners were specifically less likely to generate the correct response from the outset. The findings lend credence to the role of a developmental automatization process and indicate that developmental improvements in reasoning accuracy are at least partially driven by an improvement in the accuracy of our intuitions.

Keywords: Dual-process theory; Intuition; Development; Two-response paradigm

1 Introduction

Although humans have unique capacities to reason, we do not always reason correctly. Decades of reasoning and decision-making research has documented that people readily violate the most elementary logical principles (e.g., Kahneman, 2011). Popular dual process models of thinking have attributed this bias to the human tendency to rely on fast and effortless intuitive thinking. Although this intuitive processing can sometimes be useful, it can also cue salient "heuristic" responses (i.e., responses based on background beliefs or stereotypes) that conflict with more logical principles. It is generally believed that sound reasoning in these cases requires us to engage in more effortful deliberate reasoning and correct our intuitions (Evans & Stanovich, 2013; Kahneman, 2011).

Consider for example the following syllogism: "All people who live in The White House live in America. Joe Biden lives in America. Therefore, Joe Biden lives in the White House". Logically speaking, the conclusion is not valid (i.e., All X are Y, does not imply that all Y are also X). However, many people will readily accept the invalid conclusion simply because it is believable. Hence, by intuitively relying on the conclusion believability rather than on its logical structure, reasoners end up being biased. The idea is that avoiding such biased reasoning requires us to engage in more deliberate, logical thinking and override the initial belief-based response. However, because human cognitive misers typically tend to minimize effortful deliberations, they will often stick to the intuitively cued response and err.

Although this traditional dual process conceptualization has been highly influential, its key underlying processing assumptions have come under fire (e.g., Bago & De Neys, 2017; De Neys, 2017; Handley & Trippas, 2015; Pennycook, Fugelsang, & Koehler, 2015; Reyna, 2012; Thompson et al., 2018). Consider, for example, findings with the two-response paradigm (Thompson, Prowse Turner, & Pennycook, 2011). In this paradigm reasoners are presented with classic reasoning problems and are asked to give two consecutive responses. First, they have to answer as fast as possible with the first intuitive response that comes to mind. Next, they can take all the time they want to reflect on the problem and give a final response. To make sure that the initial response is generated intuitively it often has to be generated under concurrent secondary task load and/or time-pressure. The rationale here is that because deliberation is known to be time and cognitive resource demanding, one can minimize possible deliberation by depriving reasoners from these re-

sources. Interestingly, the two-response results indicate that sound reasoners typically already generate the correct response in the initial, intuitive response stage (e.g., Bago & De Neys, 2017, 2019; Raoelison & De Neys, 2019; Thompson & Johnson, 2014). Hence, pace the traditional dual process view, correct logical reasoners do not need to deliberate to correct erroneous intuitions; they rather generate the correct response intuitively.

Based on this and related findings various scholars have postulated that elementary logical processing can also be done intuitively (e.g., De Neys & Pennycook, 2019, for review). However, the suggestion that people can do "fast logic" or have "logical intuitions" has not been uncontroversial (e.g., Aczel, Szollosi, & Bago, 2016; Ferreira et al., 2016; Klauer & Singmann, 2013; Singmann, Klauer, & Kellen, 2014). Indeed, this goes counter to one of the key assumptions in traditional theories of reasoning and decision-making, namely that logical reasoning is typically slow and effortful (e.g., Kahneman, 2011). As potential implications can be far-reaching, even proponents have acknowledged that further validation is needed (De Neys, 2017; De Neys & Pennycook, 2019).

A key question concerns the origin of people's intuitive logical processing. It has been suggested that logical intuitions arise from an automatization process (Bago & De Neys, 2020; De Neys, 2012; Evans, 2019; Stanovich, 2018). Many of the key elementary logico-mathematical principles that are evoked in classic reasoning tasks are taught during the (high) school curriculum. The basic idea is that the year-long repeated exposure to these principles over the school years allowed reasoners to practice them to automaticity (De Neys, 2012). That is, sound adult reasoners might have instantiated the critical logical knowledge structures or *mindware* to such as degree that they can be applied without any deliberation (Stanovich, 2018).

In theory, the automatization assumption is not unreasonable. The underlying mechanism of a controlled-to-automatic processing transfer is a hallmark of basic information processing theories in cognitive science (e.g., Shiffrin & Schneider, 1977). The problem is that the assumption has not been directly tested in the case of logical reasoning (De Neys & Pennycook, 2019). The present study presents a simple but critical test. If the contested intuitive logical responding in classic reasoning tasks among adult participants results from an automatization process, then correct intuitive responding should be less likely among younger reasoners who have had—by definition—less opportunity to learn and automatize them.

To verify this prediction, we tested a group of younger and older reasoners at the start

(7^{th} grade) and end (12^{th} grade) of secondary education with a two-response paradigm. We reasoned that near the end of secondary education, the hypothesized automatization process should be largely completed and reach near adult levels. We opted for the 7^{th} graders as youngest contrast group to side-step possible complications with respect to material familiarity and reading/task demands among even younger reasoners (see further). We opted for 12^{th} graders (rather than adults) as oldest contrast group to make sure that both groups were tested in a similar (school) test context. To get an indication of the robustness of the findings, participants were presented with two different classic "bias" tasks (i.e., belief-bias syllogisms and base-rate neglect task, see further) in which a cued heuristic response conflicted with the correct response and a set of control problems. For the oldest group of 12^{th} graders we expected that just as with adults, within a single trial, correct responses in the final response stage would typically be preceded by a correct initial response (i.e., an initial-and-final correct response). Key prediction was that this intuitive correct responding would be less likely for the younger group of 7^{th} graders.

To avoid confusion, there is little discussion about the fact that children's reasoning becomes more accurate and less biased with age (Houdé, 2014; Markovits & Barrouillet, 2004; but see also Davidson, 1995; Markovits et al., 2019; Reyna & Ellis, 1994). Our point concerns the nature of this more accurate reasoning. Dual process theorists have traditionally attributed the developmental performance improvement to an increase in deliberate processing: An age-related increase in cognitive capacity will make it more likely that older reasoners manage to complete the effortful correction of erroneous intuitions (e.g., Barrouillet, 2011; De Neys & Everaerts, 2008; De Neys & Van Gelder, 2009; Houdé & Borst, 2015; Kokis et al., 2002). The critical prediction of the logical intuition view is that any age-related decrease in biased reasoning should be at least partially driven by an increase in accurate initial intuitive responses. Our developmental two-response study allowed us to test this prediction directly.

2 Methods

2.1 Preregistration

The study design and hypothesis were preregistered on the Open science Framework (https://osf.io/5envf). No specific analyses were preregistered.

2.2 Participants

We recruited 83[1] French pupils in 7th grade (43 female, Mean age = 12.4 years, *SD* = 0.5 years), and 95 students in 12th grade (49 female, Mean age = 16.8 years, *SD* = 0.6 years). A power analysis based on directional, one-sided t-tests had indicated that a sample size of 51 would allow us to detect a medium effect size with power of .80. We thus preregistered 51 subjects per age group but we decided to include all subjects whose parents or legal guardians gave their consent (consent forms were returned on the day of the study).

2.3 Material

This experiment adopted two classic tasks that have been widely used to study biased reasoning. For each of those, participants had to solve four standard, "conflict" problems, four control, "no-conflict" problems, and two control neutral problems (see further).

Base Rate (BR). Participants solved a total of ten base-rate problems adapted in French from Pennycook et al. (2014). Participants always received a description of the composition of a sample (e.g., "This study contained I.T engineers and professional boxers"), base rate information (e.g., "There were 995 engineers and 5 professional boxers") and a description that was designed to cue a stereotypical association (e.g. "This person is strong"). Participants' task was to indicate to which group the person most likely belonged. The problem presentation format was based on Pennycook et al. (2014). The base rates and descriptive information were presented serially and the amount of text that was presented on screen was minimized. First, participants received the names of the two groups in the sample (e.g., "This study contains clowns and accountants"). Next, under the first sentence (which stayed on the screen) we presented the descriptive information (e.g., Person 'L' is funny). The descriptive information specified a neutral name (Person 'L') and a single word personality trait (e.g., "strong" or "funny") that was designed to trigger the stereotypical association. Finally, participants received the base rate probabilities. The following illustrates the full problem format:

This study contains clowns and accountants.

[1]Data from one 7th grader was removed as we were informed after the experiment that he wasn't fluent in French.

Person 'L' is funny.

There are 995 clowns and 5 accountants.

Is Person 'L' more likely to be:
-A clown
-An accountant

Four of the presented problems were conflict items, four other were no-conflict items, and the remaining two were neutral items. In conflict items the base rate probabilities and the stereotypical information cued opposite responses, while in no-conflict items they cued the same response. No-conflict items were used to ensure that participants were focusing on the task throughout the experiment and would not be random-guessing. The description in neutral items cued an association that applied equally to both groups. Neutral items could therefore be used to assess whether participants knew the probability principle underlying the task (Frey, Johnson, & De Neys, 2018). Two sets of items were used for counterbalancing purposes. The same contents were used but the conflict and no-conflict status of the items in each set was reversed by switching the base rates of the two groups.

Each item started with the presentation of a fixation cross for 1000 ms. After the fixation cross disappeared, the sentence which specified the two groups appeared for 2000 ms. Then the stereotypical information appeared, for another 2000 ms, while the first sentence remained on the screen. Finally, the last sentence specifying the base rates appeared together with the question and two response alternatives. Once the base-rates and question were presented participants were able to select their answer by clicking on it. Conflict and no-conflict items were presented in random order, followed by the neutral items. Neutral items were presented last to avoid cueing of the correct response or interference on the other items.

Syllogism (SYL). We used the same syllogistic reasoning task as Bago & De Neys (2017). Participants were given ten syllogistic reasoning problems based on Markovits & Nantel (1989) and adapted in French. Each problem included a major premise (e.g., "All dogs have four legs"), a minor premise (e.g., "Puppies are dogs"), and a conclusion (e.g., "Puppies have four legs"). Participants were instructed that they had to accept that the

premises were true and their task was to evaluate whether the conclusion followed logically from the premises. We used the following format:

All dogs have four legs

Puppies are dogs

Puppies have four legs

Does the conclusion follow logically?
-Yes
-No

In four of the items the believability and the validity of the conclusion conflicted (conflict items, two problems with an unbelievable–valid conclusion, and two problems with a believable–invalid conclusion). In four other items the logical validity of the conclusion was in accordance with its believability (no-conflict items, two problems with a believable–valid conclusion, and two problems with an unbelievable–invalid conclusion). Lastly, two neutral items used abstract content with the same logical structures as the other items (e.g., "All X are Y..."). Conflict and no-conflcit items were presented in random order, followed by the neutral items.

The premises and conclusion were presented serially. Each trial started with the presentation of a fixation cross for 1000 ms. After the fixation cross disappeared, the first sentence (i.e., the major premise) was presented for 2000 ms. Next, the second sentence (i.e., minor premise) was presented under the first premise for 2000 ms. After this interval was over, the conclusion together with the question "Does the conclusion follow logically?" and two response options (yes/no) were presented right under the premises. Once the conclusion and question were presented, participants could give their answer by clicking on it.

Rating task. Our item material was based on a pilot study (see further) to make sure it was familiar to younger reasoners. Following De Neys & Vanderputte (2011), we also asked participants in our main reasoning study to rate the material at the end of the study to double-check that it cued the intended stereotypes and beliefs among both age groups. Participants were asked to indicate to which degree they agreed with statements on a scale ranging from -5 to +5. For syllogisms, the eight conclusions from the conflict

and no-conflict problems were presented as separate statements to be evaluated. For base-rate items, two statements were presented at the same time, which were associating each group with the adjective (e.g., "A clown is funny." and "An accountant is funny."). Eight pairs of statements were rated, corresponding to the eight conflict and no-conflict items in the reasoning task. Participants were presented with the material from only one task for rating, which was randomly selected for each participant. Results established that the material worked as intended. In both age groups, believable conclusions and intended stereotypical associations were consistently rated higher than unbelievable conclusions and the contrasting association, with minimal variability between the age groups (for a detailed overview, see Supplementary material, Figure S1).

Two-response format. We used the two-response paradigm (Thompson, Prowse Turner, & Pennycook, 2011) to elicit both an initial, intuitive response and a final, deliberate one. Participants had to provide two answers consecutively to each reasoning problem. To minimize the possibility that deliberation was involved in producing the initial response, participants had to provide their initial answer within a strict time limit while performing a concurrent load task (see Bago & De Neys, 2017, 2019; Raoelison & De Neys, 2019). The load task was based on the dot memorization task (Miyake et al., 2001) as it had been successfully used to burden executive resources during reasoning (e.g., De Neys, 2006; Franssens & Neys, 2009). Participants had to memorize a complex visual pattern (i.e., 4 crosses in a 3x3 grid) presented briefly before each reasoning problem. After answering the reasoning problem the first time (i.e., intuitively), participants were shown four different patterns (i.e., with different cross placings) and had to identify the one presented earlier.

The precise initial response deadline for each task, set at 3 seconds, was based on the pretesting by Bago & De Neys (2017, 2019). The allotted time corresponded to the time needed to simply read the problem conclusion, question, and answer alternatives (i.e., the last part of the serially presented problem) in each task and move the mouse. The load and deadline were applied only during the initial response stage and not during the subsequent final response stage in which participants were allowed to deliberate (see further).

2.4 Procedure

Pilot experiment. The two-response paradigm can be quite challenging. Although seventh graders can be expected to be proficient readers, we ran a pilot experiment to verify that our youngest participants could still read and understand the problems under the deadline and load burden. A convenience sample of 13 participants (Mean age = 12.4, $SD = 0.7$) were recruited through personal networks.

When presented with the reasoning problems under two-response conditions, they missed the deadline on 14 (10.8%) syllogisms and 17 (13.1%) base-rate problems. Similarly, they failed the load memorization task for 19 (14.6%) syllogisms and 23 (17.7%) base-rate problems. In combination, they missed 29 (22.3%)trials for syllogisms and 37 (28.5%) trials for base-rate problems, indicating that both the deadline and the load memorization task were challenging but could be met on the vast majority of trials.

We checked problem understanding by analyzing the performance on the no-conflict problems. Recall that on these problems mere intuitive reasoning is cuing the correct response. If participants can properly read the problem under two-response burden, they should show high (near ceiling) accuracy. Results showed that performance on the remaining (i.e., when response deadline and load memorization were both met) no-conflict trials was indeed very high for both the base-rate task(initial accuracy: $M = 91.6\%$, $SD = 16.7$; final accuracy: $M = 94.2\%$, $SD = 15$) and the syllogisms (initial accuracy: $M = 87.2\%$, $SD = 17.9$; final accuracy: $M = 96.2\%$, $SD = 13.9$), indicating that participants were able to read and solve the problems in spite of the deadline and load memorization task.

Note that after the reasoning task our pilot study participants were also presented with the material rating task (see above). The pilot rating task included additional items and we selected the items with the highest familiarity for the main study.

Main experiment. The experiment was run individually on iPads. The participants were tested in their classrooms in groups under the supervision of at least two experimenters.

Upon starting the experiment, participants were told that the study would take about thirty minutes. They were then presented with the two reasoning tasks in a random order. Before the first task, they received the following general information:

Please read these instructions carefully!

The following task is composed of 8 questions and a couple of practice questions. It should take about 10 minutes to complete and it demands your full attention.

In this task we'll present you with a set of reasoning problems. We want to know what your initial, intuitive response to these problems is and how you respond after you have thought about the problem for some more time.

Hence, as soon as the problem is presented, we will ask you to enter your initial response. We want you to respond with the very first answer that comes to mind. You don't need to think about it. Just give the first answer that intuitively comes to mind as quickly as possible.

Next, the problem will be presented again and you can take all the time you want to actively reflect on it. Once you have made up your mind you enter your final response. You will have as much time as you need to indicate your response.

In sum, it is really crucial that you give your initial response as fast as possible. Afterwards, you can take as much time as you want to reflect on the problem and select your final response.

Click on Next to proceed to the task instructions.

Next, the first reasoning task was presented. The presentation format was explained and the deadline for the initial response was introduced. Participants solved two unrelated practice reasoning problems to familiarize themselves with the deadline and the procedure. Next, they solved two practice matrix recall problems (without concurrent reasoning problem). Finally, at the end of the practice, they had to solve the two earlier practice reasoning problems under cognitive load.

Every task trial started with a fixation cross shown for 1 second. Next, the target pattern for the memorization task was presented for 2 seconds. We then presented the problem introduction (i.e., group composition for the base-rate items, or major premise of the syllogisms, see Material) for 2 seconds. The second sentence (i.e., adjective of the base-rate item, or the minor premise of the syllogisms) was then displayed under the introduction for 2 seconds as well. Afterward the full problem was presented and

participants were asked to enter their initial response. The initial response deadline was 3 seconds for all problems (see Material). One second before the deadline the background of the screen turned yellow to warn participants about the upcoming deadline. If they did not provide an answer before the deadline, they were asked to pay attention to providing an answer within the deadline on subsequent trials.

After the initial response was entered, participants were presented with four matrix patterns from which they had to choose the correct, to-be-memorized pattern. Once they provided their memorization answer, they received feedback as to whether it was correct. If the answer was not correct, they were also asked to pay more attention to memorizing the correct dot pattern on subsequent trials.

Finally, the full item was presented again, and participants were asked to provide a final response.

The color of the answer options was green during the first response, and blue during the final response phase, to visually remind participants which question they were answering. Therefore, right under the question we also presented a reminder sentence: "Please indicate your very first, intuitive answer!" and "Please give your final answer.", respectively, which was also colored as the answer options.

After completing the first task, a short transition indicated the overall progress (e.g., "You've completed the first task! Click on Next when you are ready to start the second task."). The second reasoning task was then introduced.

After participants completed the second reasoning task, they were asked to rate the material.

Finally, participants completed a page with demographic questions.

2.5 Exclusion criteria

As in previous studies (e.g., Bago & De Neys, 2017; Raoelison & De Neys, 2019), We discarded trials where participants failed to provide a response before the deadline or failed the load memorization task because we could not guarantee that the initial response for these trials did not involve deliberation.

For the base-rate neglect task, older participants missed the deadline on 28 (2.9%) trials and further failed the load task on 119 (12.5%) trials, leading to 803 remaining trials out of 950 (84.5%). On average they each contributed 3.3 (SD = 0.7) standard

conflict problem trials, 3.4 (SD = 0.6) control no-conflict trials, and 1.8 (SD = 0.4) neutral trials. Younger participants missed the deadline on 92 (11.1%) trials and further failed the load task on 132 (15.9%) trials, leading to 606 remaining trials out of 830 (73%). On average they each contributed 3 (SD = 1.1) standard problem trials, 3 (SD = 1.1) control no-conflict trials, and 1.3 (SD = 0.7) neutral trials. Altogether, we kept 1409 trials out of 1780 (79.2%) for the base-rate task.

For syllogisms, older participants missed the deadline on 49 (5.2%) trials and further failed the load task on 66 (6.9%) trials, leading to 835 remaining trials out of 950 (87.9%). On average they each contributed 3.5 (SD = 0.7) standard conflict problem trials, 3.6 (SD = 0.6) control no-conflict trials, and 1.7 (SD = 0.5) neutral trials. Younger participants missed the deadline on 75 (9%) trials and further failed the load task on 127 (15.3%) trials, leading to 628 remaining trials out of 830 (75.7%). On average they each contributed 3 (SD = 1) standard problem trials, 3.1 (SD = 1) control no-conflict trials, and 1.6 (SD = 0.6) neutral trials. Altogether, we kept 1463 trials out of 1780 (82.2%) for syllogisms.

3 Results and discussion

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

For consistency with previous work we first analyze the accuracy data and then move on to the critical direction of change analysis.

3.1 Accuracy

Base rate. Visual inspection of Figure 1 indicates that older participants outperformed younger participants on conflict items, both at the initial and the final response stage. In addition, the older participants tended to improve from the initial to the final response stage but not younger participants. We ran a 2 (age group, young or old) x 2 (response stage, initial or final) mixed ANOVA on conflict problem accuracy with age group as a between-subject factor and response stage as a within-subject factor to test these trends. It revealed a main effect of age group, $F(1, 173) = 77.01$, $p < .001$, $\eta_p^2 = .31$, a main effect

of response stage, $F(1, 173) = 21.78$, $p < .001$, $\eta_p^2 = .11$, and a significant interaction, $F(1, 173) = 26.91$, $p < .001$, $\eta_p^2 = .14$. Paired comparisons showed that conflict accuracy for the older group significantly increased from the initial ($M = 51.1$, $SD = 32.2$) to the final ($M = 72.8$, $SD = 30.4$) response stage, $t(94) = 6.32$, $p < .001$, $r = .55$, but this wasn't the case for the younger group (initial: $M = 17.6$, $SD = 28.2$; final: $M = 16.5$, $SD = 30.3$), $t(79) = -0.45$, $p = .652$, $r = .05$. Hence, consistent with previous two-response studies with adults, deliberation in the final response stage increased accuracy for our 12^{th} graders. Interestingly, younger 7^{th} graders did not yet show this deliberation benefit.

As Figure 1 further indicates, for the control no-conflict problems we get overall very high accuracy rates. For completeness, we also ran a 2 (age group, young or old) x 2 (response stage, initial or final) mixed ANOVA on no-conflict problem accuracy. Results pointed to a main effect of age group, $F(1, 173) = 12.85$, $p < .001$, $\eta_p^2 = .07$. The main effect of response stage was not significant, $F(1, 173) = 2.76$, $p = .099$, $\eta_p^2 = .02$, neither was the interaction, $F(1, 173) = 1.68$, $p = .197$, $\eta_p^2 = .01$. The age effect indicates that while both groups performed near ceiling, older participants were slightly more accurate (initial: $M = 96.1$, $SD = 23.3$; final: $M = 95.5$, $SD = 17.2$) than younger participants (initial: $M = 89.4$, $SD = 31.2$; final: $M = 85.1$, $SD = 32.1$) on no-conflict items. Nevertheless, consistent with our pilot study, the high initial accuracy rates confirm that (younger) participants could read and process the material and refrained from random guessing.

Finally, Figure 1 also shows the accuracy results for neutral items. A mixed ANOVA revealed a main effect of age group, $F(1, 167) = 45.53$, $p < .001$, $\eta_p^2 = .21$, and response stage, $F(1, 167) = 14.36$, $p < .001$, $\eta_p^2 = .08$, but no significant interaction, $F(1, 167) = 3.27$, $p = .072$, $\eta_p^2 = .02$. While older participants (initial: $M = 86.8$, $SD = 22.7$; final: $M = 92.1$, $SD = 21.9$) overall outperformed younger ones (initial: $M = 50$, $SD = 35.8$; final: $M = 64.9$, $SD = 34.5$), as with the conflict problems, it seems that both age groups improved from initial to final response stage on the neutral items. Hence, even younger participants were able to improve on neutral items when allowed to deliberate.

Syllogisms. Figure 1 (bottom panel) also shows the accuracy results for the syllogisms. As the figure indicates, although accuracy rates are overall lower, we observe very similar trends as with the base-rate problems. A 2 (age group, young or old) x 2 (response stage, initial or final) mixed ANOVA on conflict accuracy revealed a main effect of age group, $F(1, 172) = 4.61$, $p = .033$, $\eta_p^2 = .03$, a main effect of response stage, $F(1, 172) = 7.8$, p

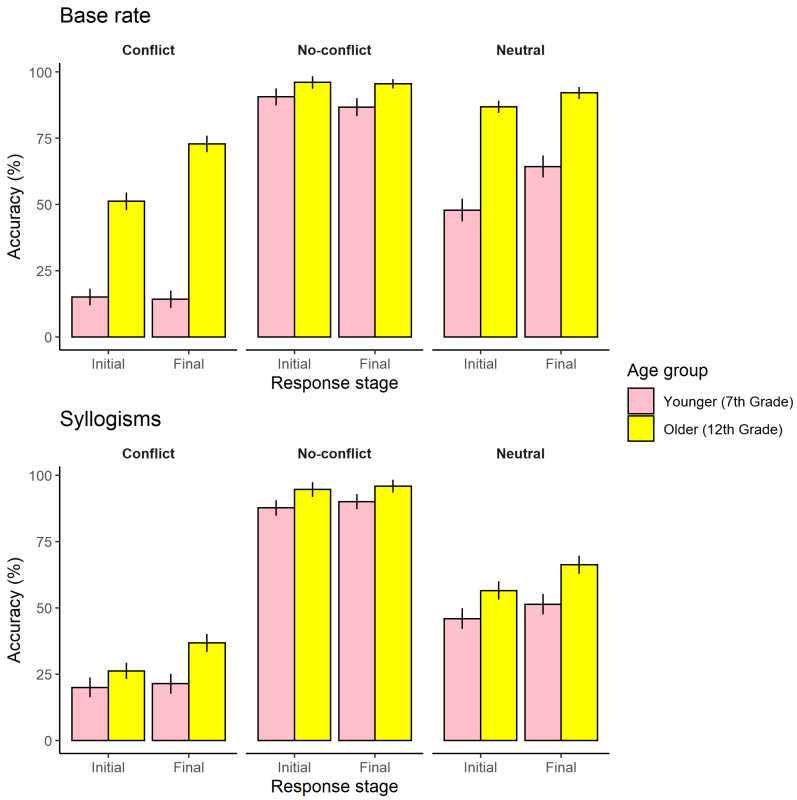

Figure 1. Accuracy for base-rate problems and syllogisms separated by problem type. *Note.* Error bars represent standard error.

= .006, η_p^2 = .04, and a significant interaction, $F(1, 172) = 4.62$, $p = .033$, $\eta_p^2 = .03$. As was the case for the base-rate task, pairwise comparisons indicated that older participants improved from the initial to the final response stage (initial: $M = 26.2$, $SD = 29.5$; final: $M = 36.8$, $SD = 32.7$), $t(94) = 3.85$, $p < .001$, $r = .37$, but younger participants didn't (initial: $M = 21$, $SD = 34.9$; final: $M = 22.4$, $SD = 34.3$), $t(78) = 0.41$, $p = .681$, $r = .05$.

A mixed ANOVA on no-conflict accuracy revealed a significant main effect of age group, $F(1, 173) = 10.16$, $p = .002$, $\eta_p^2 = .06$. The main effect of response stage was not significant, $F(1, 173) = 1.49$, $p = .224$, $\eta_p^2 = .01$, neither was the interaction, $F(1, 173) = 0.11$, $p = .742$, $\eta_p^2 < .01$. As with the base-rate problems, although older participants (initial: $M = 94.6$, $SD = 26.9$; final: $M = 96$, $SD = 23.9$) scored slightly higher than younger ones (initial: $M = 88$, $SD = 26.3$; final: $M = 90.3$, $SD = 26.1$), both groups

performed near ceiling on no-conflict items.

Finally, a mixed ANOVA on the neutral item accuracy rates found a main effect of age group, $F(1, 167) = 6.5$, $p = .012$, $\eta_p^2 = .04$, and response stage, $F(1, 167) = 7.33$, $p = .007$, $\eta_p^2 = .04$. The interaction was not significant, $F(1, 167) = 0.3$, $p = .585$, $\eta_p^2 = .02$. Hence, as with the base-rate problems, older participants (initial: $M = 56.5$, $SD = 33.5$; final: $M = 66.3$, $SD = 32.2$) performed better than the younger ones (initial: M = 45.5, SD = 33.1; final: M = 51.9, SD = 33.1)but even younger participants tended to benefit from additional deliberation when solving neutral problems.

3.2 Direction of change

To have a deeper look at how much participants changed their initial answer after deliberation–and whether correct final responses were already generated intuitively–we ran a direction-of-change analysis (Bago & De Neys, 2017). Since participants are asked to provide two responses on each trial, this results in four possible direction of change categories: *00* (both initial and final responses incorrect), *01* (incorrect initial response but correct final response), *10* (correct initial response but incorrect final response) and *11* (both initial and final responses correct). For brevity, we focus on the critical conflict items and neutral items (as expected, no conflict items showed predominantly *11* responses throughout, see Supplementary Material). We calculated the proportion of each direction category in each task for every individual.[2] Figure 2 shows the results.

3.2.1 Conflict items

Base rate. As Figure 2 shows, the main pattern seems to be a clear drop in *00* with age that is specifically accompanied by an increase in *11* and *01* responses. Directional t tests showed that the increase in *11* (young: $M = 11\%$, $SD = 26.4$; old: $M = 49.1\%$, $SD = 42.9$; 38.1% more), $t(159.31) = 7.18$, $p < .001$ (one-tailed), $r = .49$, and *01* trials (young: $M = 5.4\%$, $SD = 15.8$; old: $M = 23.7\%$, $SD = 32.1$; 18.3% more), $t(141.77) = 4.9$, $p < .001$ (one-tailed), $r = .38$, were both significant. Hence, in line with classic developmental claims, older reasoners were better able to deliberately correct an intuitive response in

[2]This reflects how likely an individual is to show each specific direction of change pattern. Thus, for any individual, P(*00*) + P(*01*) + P(*10*) + P(*11*) = 1.

case it was erroneous. But critically, consistent with our hypothesis, older reasoners were also far more likely to generate a correct intuitive response from the outset.

For exploratory purposes, we also directly contrasted the difference between proportions for the *11* and *01* trials (i.e., %*11*-%*01*) in the two age groups. This gives us an indication of how much more likely a correct response was generated intuitively rather than after correcting an initial erroneous response. This difference was larger in the older group ($M = 25.4\%$, $SD = 64.5\%$) than in the younger group ($M = 5.6\%$, $SD = 30.3\%$). This suggests that the increase in accuracy observed with age relies more on better intuitions than correction. We come back to this issue in the discussion but note however that values are positive in both age groups. As Figure 2 shows, even the youngest participants were more likely to produce a *11* trial than a *01* trial. Hence, although younger reasoners are overall especially less likely to generate correct intuitive responses, the few younger reasoners who do give correct responses typically are also more likely to generate these intuitively. The point is that in absolute terms this is far less likely than among older reasoners.

Syllogisms. As Figure 2 shows, although the trends were somewhat less pronounced than with the base-rate problems, we observed a similar pattern on the syllogisms with an age-related increase in *11* and *01* responses. The proportion of *11* trials was higher for older participants ($M = 22.2\%$, $SD = 31.7$) than for younger participants ($M = 14\%$, $SD = 27.2$), $t(171.82) = 1.83$, $p = .034$ (one-tailed), $r = .14$. Likewise, the proportion of *01* trials was higher for older participant ($M = 14.6\%$, $SD = 23.2$) than for younger participants ($M = 8.3\%$, $SD = 21.2$), $t(170.46) = 1.85$, $p = .033$ (one-tailed), $r = .14$. As was the case for base-rate problems, the contrast between *11* and *01* proportions was larger for older participants ($M = 7.6\%$, $SD = 42$) than for younger participants ($M = 5.7\%$, $SD = 36$), although he difference was small.

3.2.2 Neutral items

Base rate. As Figure 2 shows, there was a clear increase in *11* responding with age (old: $M = 83.7\%$, $SD = 33$; young: $M = 42.6\%$, $SD = 45.1$; 41.1% difference), $t(129.09) = 6.59$, $p < .001$ (one-tailed), $r = .50$. However, the figure also shows a decrease in *01* responding with age rather than an increase (old: $M = 8.4\%$, $SD = 21.5$; young: $M = 22.3\%$, $SD = 37.1$; 13.9% more). A directional t test indicated that this decrease was

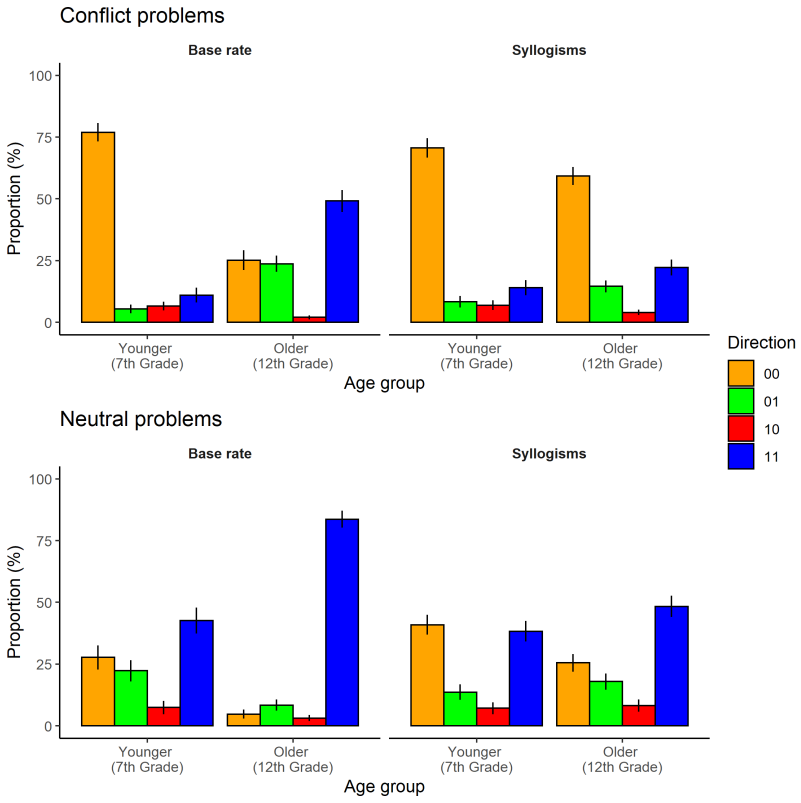

Figure 2. Proportions of Direction of Change Categories for Conflict and Neutral Problems. *Note.* Error bars represent standard error. Proportions for each direction were computed for each individual. Averages were then computed separately for each direction. *Note.* *00* = both initial and final responses incorrect, *01* = incorrect initial response but correct final response, *10* = correct initial response but incorrect final response and *11* = both initial and final responses correct.

significant, $t(110.1) = 2.86$, $p = .003$ (one-tailed), $r = .26$. That is, older participants were able to generate more intuitive correct responses than younger participants but were less likely to correct an initial erroneous response. To put this result in context, one needs to consider the high initial accuracy for neutral items, which was above 80% for older participants and below 50% for younger participants. Thus, the latter simply had a greater amount of initial responses to correct to begin with. Since solving neutral items only required applying probability principles (i.e., without an interfering cued heuristic

response), this suggests that in the cases when younger participants could not apply those principles intuitively, they could still access them and apply them when given enough time, while older participants were better able to access and apply them intuitively. As with the conflict items, the difference between *11* and *01* proportions was larger for the older group ($M = 75.3\%$, $SD = 50.5\%$) than for the younger group ($M = 20.3\%$, $SD = 70.2\%$).

Syllogisms. Older participants had a higher proportion of *11* trials ($M = 48.4\%$, $SD = 40.9\%$) than younger participants ($M = 38.3\%$, $SD = 36.2\%$), $t(166.47) = 1.7$, $p = .046$ (one-tailed). They also had a higher proportion of *01* trials ($M = 17.9\%$, $SD = 31.1\%$) than younger participants ($M = 13.6\%$, $SD = 27.7\%$) but the trend was not significant, $t(166.34) = 0.951$, $p = .172$ (one-tailed). The difference between proportions was larger for the older participants ($M = 30.4\%$, $SD = 62.4$) than for the younger participants ($M = 24.7\%$, $SD = 53.6$).

3.3 Exploratory analyses

Although this was not the focus of our study, for exploratory purposes we also recorded participants' response latencies. The interested reader can find an overview in the Supplementary Material (Figure S3). Results showed that response latencies for the initial responses (i.e., generated under time-pressure) were virtually identical in the two age groups (overall $M = 1.62$s, $SD = 0.31$s). For final response latencies the only systematic trend across the base-rate problems and syllogisms were longer conflict problem latencies for the older group (Base-rate problems: young $M = 5.77$s, $SD = 3.74$s, old $M = 6.37$s, $SD = 3.22$s. Syllogisms: young $M = 4.52$s, $SD = 2.50$s, old $M = 6.57$s, $SD = 4.18$s) which presumably reflects the more extensive deliberation in this age group.

For consistency with previous work we also looked at conflict detection effects (De Neys & Glumicic, 2008). Previous studies have shown that when adult participants fail to solve a conflict problem correctly they often show some sensitivity to their errors as reflected, for example, in longer latencies for incorrectly solved conflict problems, than for correctly solved no-conflict problems (e.g., De Neys, 2012; Frey, Johnson, & De Neys, 2018; Pennycook, Fugelsang, & Koehler, 2015). Developmental studies have shown that these conflict detection effects are less pronounced for younger reasoners (De Neys, Cromheeke, & Osman, 2011; De Neys & Feremans, 2013; Lanoë et al., 2017). The present study confirmed

these results[3] with a consistent trend towards a more pronounced conflict detection effect in the older than younger age group (see Figure S4).

4 General discussion

In the present study we contrasted the reasoning performance of younger (7^{th} grade) and older (12^{th} grade) reasoners with a two-response paradigm. Results showed that, consistent with traditional developmental claims, older reasoners were more likely to deliberately correct an initial erroneous response than younger reasoners. Critically, however, consistent with our hypotheses, older reasoners were also far more likely to generate the correct response from the outset.

This increase in correct intuitive responding with age lends credence to the suggestion that correct intuitive responding among adults might result from an automatization process (De Neys, 2012). As students are repeatedly exposed to logico-mathematical principles throughout their school curriculum they have the opportunity to practice them to automaticity. With enough repetition, this *mindware* might become so instantiated that it can be applied effortlessly, without deliberation (Stanovich, 2018).

Consequently, correct intuitive responding will be less likely among younger reasoners who have had less opportunity to practice and automatize the principles.

Our findings can help to inform the theoretical debate concerning the nature of people's alleged intuitive logical competence. However, the results have also interesting implications for developmental psychologists. As we noted in the introduction, traditionally the reasoning community has tended to attribute age-related reasoning improvements to increased deliberate correction (e.g., Barrouillet, 2011; De Neys & Everaerts, 2008; Houdé & Borst, 2015; Kokis et al., 2002). Although we observed such improved correction, the age-related decrease in "biased" responding was specifically accompanied by an increase in correct intuitive responses. Hence, older reasoners are not necessarily less biased because they are better able to correct their intuitions when they are wrong but rather because their intuitions are more accurate. Interestingly, the developmental work of Reyna and colleagues (e.g., Reyna, 2012; Reyna & Brainerd, 2011) has long argued for such an alter-

[3] As suggested in previous two-response studies (Bago & De Neys, 2017; Thompson & Johnson, 2014), we focused on final latencies for conflict detection as initial latencies are not reliable due to the deadline (e.g., Janssen, Raoelison, & de Neys, 2020).

native view. Their fuzzy-trace theory has put intuitive processing at the apex of cognitive functioning. They have argued that this is mediated by a switch from more verbatim to more intuitive gist-based representations (e.g., see Reyna et al., 2017). Although the two-response findings are agnostic about the underlying representations, they lend credence to the general point that reasoning development should not be conceived as a mere switch from intuitive to deliberate processing per se (see also Markovits et al., 2019). Learning to accurately intuit seems to be at least as important.

Our key result concerning the age-related increase in intuitive correct responding was observed for traditional conflict problems in which a cured heuristic response conflicted with the correct response and on neutral control problems in which no heuristic response was cued. Interestingly, our results also pointed to a difference between the two problem types. On conflict problems, younger reasoners' overall accuracy did not increase after deliberation. Younger reasoners did benefit from additional deliberation, however, on the neutral problems. The key difference between the conflict and neutral problems is that reasoners are not being faced with an interfering intuitive heuristic response on the neutral problems. This increased neutral problem performance suggests that biased younger reasoners do know the critical logical principles to some extent and can apply them when given the time to deliberate (e.g., De Neys & Van Gelder, 2009). However, their deliberation is specifically less effective when they need to deal with a competing salient heuristic response.

We also need to consider possible methodological reservations. First, it can be argued that the poorer performance of younger reasoners results from a reading or guessing confound. In theory, it could be that even younger reasoners have accurate logical intuitions but that they simply failed to read and understand the problem material under time-pressure and load. As we noted, our pretest and control no-conflict problems findings argue against such a confound. Although there was a slight age-related performance increase on the no-conflict problems, even our youngest 7^{th} graders showed near ceiling performance and gave intuitive correct responses on the vast majority of no-conflict trials. If the two-response constraints would have interfered with 7^{th} graders' basic problem comprehension per se, they should not have performed so well on the no-conflict problems.

Alternatively, one could also argue that our two-response paradigm might not have been demanding enough for the oldest participants and still allowed them to engage in deliberation at the initial response stage. However, here it should be noted that the

specific two-response paradigm we used has been extensively validated with adults. It combines instruction, time-pressure, and load manipulations that have all been shown to minimize deliberation in isolation (Bago & De Neys, 2017). Hence, even for adults possible deliberation is extremely minimized (Bago & De Neys, 2020). Still, one may argue that, given that older reasoners will have more cognitive resources, our experimental constraints were less demanding for older reasoners than younger ones. Hence, in theory, older reasoners might have had at least more opportunity to engage in deliberation during the initial response stage, which could have driven the higher initial accuracy. Given that deliberation is more time-demanding than intuitive processing, if it was indeed the case that older reasoners were deliberating in the initial response stage, we should also have observed longer initial response times for the deliberating older reasoners than for the intuiting younger reasoners. However, we found no evidence for such a systematic effect (overall initial response time for young = 1.65s, SD = 0.33s; old = 1.59s, SD = 0.29s, see Figure S3). Taken together, these arguments argue against a systematic deliberation confound. That being said, we readily acknowledge that one can never be completely sure that a paradigm excludes all possible deliberation (e.g., see Bago & De Neys, 2019; Raoelison, Thompson, & De Neys, 2020) [4].

Another possible reservation concerns the nature of the developmental effect. For example, in theory, it is possible that older reasoners simply no longer process the heuristic information (e.g., they no longer attend to the descriptive information in the base-rate problem). In this case, they would not reason "more logically" per se but rather "less heuristically" (i.e., have weaker heuristic intuitions rather than stronger logical ones). However, our data argues against such an account. First, performance on the neutral problems–which do not cue a heuristic response–was consistently higher than on conflict problems (see Figure 1). If older reasoners would no longer process the heuristic information, they should perform equally well on conflict and neutral problems. Second, our explorative conflict detection data indicated that older reasoners showed a more pronounced conflict sensitivity than younger ones (see Figure S4) when solving conflict vs no-conflict problems. This also indicates that older reasoners are not merely disregarding heuristic cues per se.

[4]The general problem is that "fast-and-slow" dual process theories are underspecified. It is posited that deliberation is slower and more demanding than intuitive processing, but the theory does not present an unequivocal a priori criterion that allows us to classify a process as intuitive or deliberate (e.g., takes at least x time, or x amount of load, see Bago & De Neys, 2019, for an extensive discussion).

A related point concerns the possible role of developmental differences in elementary conflict resolution processes. In theory, it could be that older and younger reasoners have equally strong (i.e., automatized) logical intuitions but simply differ in how well they can intuitively resolve conflict between a heuristic and logical intuition. Although we do not object to the suggestion that older reasoners have better conflict resolution capacities, our results on the neutral problems underscore the role of better instantiated intuitive logical knowledge per se. Given that neutral problems do not cue a heuristic response, conflict resolution is not (or far less) at play. Hence, one's intuitive performance on these problems should primarily reflect the extent of one's mindware instantiation (Stanovich, 2018). As our results indicated (see Figure 1), although performance on the neutral problems was overall higher than on conflict problems, the observed developmental accuracy difference on the initial neutral problem responses in our study (base-rates +39%, syllogisms +11% for oldest group) was as pronounced as what we observed for the initial conflict problem response (base-rates +36%, syllogisms +6% for oldest group). This supports the claim that older reasoners have stronger (more automatized) logical intuitions.

Our critical trends were highly similar across the two different reasoning tasks (base-rate problems and syllogisms) we adopted. This lends credence to the generality of our findings. Nevertheless, one needs to bear in mind that the study is the first to adopt a two-response paradigm in a developmental study. Although we be believe that our pretesting and study design minimized systematic confounds, we readily acknowledge that further validation of the findings is welcome. Indeed, we hope that the study can serve as a methodological proof-of-principle that points to the potential of the two-response paradigm for future developmental work.

A general limitation of the developmental approach we took in the present study is obviously that it is essentially correlational in nature. The developmental trends we observed lend credence to the automatization hypothesis, but we did not study the automatization process "in vivo". For example, our results do not inform us about the actual timescale of the hypothesized automatization. What amount of training is needed to intuit the application of logical rules? The literature on expertise indicates that automatizing skills to expert levels can take several thousand hours of focused practice (e.g., Ericsson et al., 2006). On the other hand, the automatization literature in the cognitive sciences suggests that automatization of simple tasks can occur within a single training session (Moors & De Houwer, 2006; Shiffrin & Schneider, 1977). Clearly, although our age groups dif-

fered by 5 years, the school curriculum is not exclusively devoted to the acquisition of logico-mathematical knowledge. In addition, the "training" will often be indirect. That is, although the school curriculum familiarizes students with the general underlying principles that are at play in our reasoning tasks (e.g., impact of base-rates on likelihood estimation and conditional implication) they are not trained with these specific tasks (i.e., base-rate neglect problems and belief bias syllogisms) as such.

In this respect it might be interesting to complement the present developmental approach with an intervention or training approach in which reasoners get to directly practice the precise logical principles and reasoning task in question. We could then measure the training impact on their intuitive reasoning performance with a two-response paradigm. This could allow us, for example, to test directly how much training is required to arrive at automatization and whether certain tasks/principles (e.g., base-rate vs syllogisms) are easier to automatize than others.

Last, but not least, it is also important to highlight individual differences. Our interest in the present study lay at the group-level. our critical finding is that younger reasoners are *on average* less likely to generate correct intuitive responses than older ones. However, this obviously does not imply that all older reasoners manage to intuit correctly. Indeed, even among our 12^{th} graders biased responding remained prevalent. Such individual differences among adult (or near adult) reasoners have been attributed to differences in the strength or instantiation of one's logical intuitons (Bago & De Neys, 2017, 2020; De Neys, 2012; Stanovich, 2018). That is, not all adults will have automatized the application of the underlying logical principles to the same extent. Hence, there will be individual differnces in the degree of mindware instantiation (Stanovich, 2018). It will be this degree of mindware instantiation that will determine whether a reasoner can intuit correctly or not (and whether they will show sensitivity to their error, e.g., De Neys, 2012).

At the other end of the spectrum, our developmental trends do not imply that younger reasoners cannot reason correctly. Although it was rare, even among our 7^{th} graders we observed correct conflict responses, and critically, in case a final response was correct it was typically already correct in the initial response stage. Hence, even correct intuitive responding is not impossible for (some) 7^{th} graders. The key point is that, it is far less likely for a 7^{th} grader than for a 12^{th} grader. In sum, by applying the two-response paradigm in a developmental setting, we managed to provide some initial insight into the origin of people's alleged logical intuitions. Results lend credence to the role of a developmental

automatization process. More generally, the findings support the view that developmental improvements in reasoning accuracy are at least partially driven by an improvement in the accuracy of our intuitions.

Acknowledgments

This research was supported by a research grant (DIAGNOR, ANR-16-CE28-0010-01) from the Agence Nationale de la Recherche, France. We thank two masters students, Agnieszka Argasinska and Morgane Velly, for their involvement in piloting the experiment and in data collection.

Open data statement

Raw data can be downloaded from our OSF page (https://osf.io/6xqgt/).

References

Aczel, B., Szollosi, A., & Bago, B. (2016). Lax monitoring versus logical intuition: The determinants of confidence in conjunction fallacy. *Thinking & Reasoning*, *22*(1), 99–117. https://doi.org/10.1080/13546783.2015.1062801

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. (2019). 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. (2020). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, 1–30. https://doi.org/10.1080/13546783.2018.1552194

Barrouillet, P. (2011). Dual-process theories and cognitive development: Advances and challenges [Special Issue: Dual-Process Theories of Cognitive Development]. *Developmental Review*, *31*(2), 79–85. https://doi.org/10.1016/j.dr.2011.07.002

Davidson, D. (1995). The representativeness heuristic and the conjunction fallacy effect in children's decision making. *Merrill-Palmer Quarterly*, *41*(3), 328–346.

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner [PMID: 16683931]. *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. (Ed.). (2017). *Dual process theory 2.0*. Routledge. https://doi.org/10.4324/9781315204550

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

De Neys, W., & Everaerts, D. (2008). Developmental trends in everyday conditional reasoning: The retrieval and inhibition interplay. *Journal of Experimental Child Psychology*, *100*(4), 252–263. https://doi.org/10.1016/j.jecp.2008.03.003

De Neys, W., & Feremans, V. (2013). Development of heuristic bias detection in elementary school. *Developmental Psychology*, *49*(2), 258–269. https://doi.org/10.1037/a0028320

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

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

De Neys, W., & Van Gelder, E. (2009). Logic and belief across the lifespan: The rise and fall of belief inhibition during syllogistic reasoning. *Developmental Science*, *12*(1), 123–130. https://doi.org/10.1111/j.1467-7687.2008.00746.x

De Neys, W., & Vanderputte, K. (2011). When less is not always more: Stereotype knowledge and reasoning development. *Developmental psychology*, *47*(2), 432–441. https://doi.org/10.1037/a0021313

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (Eds.). (2006). *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511816796

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

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

Ferreira, M., Mata, A., Donkin, C., Sherman, S., & Ihmels, M. (2016). Analytic and heuristic processes in the detection and resolution of conflict. *Memory & Cognition*, *44*. https://doi.org/10.3758/s13421-016-0618-7

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

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

Handley, S. J., & Trippas, D. (2015). Chapter two - dual processes and the interplay between knowledge and structure: A new parallel processing model. In B. H. ROSS (Ed.), *Psychology of learning and motivation* (pp. 33–58). Academic Press. https://doi.org/10.1016/bs.plm.2014.09.002

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

Houdé, O. (2014). *Le raisonnement*. Presses Universitaires de France.

Houdé, O., & Borst, G. (2015). Evidence for an inhibitory-control theory of the reasoning brain. *Frontiers in Human Neuroscience*, *9*, 148. https://doi.org/10.3389/fnhum.2015.00148

Janssen, E. M., Raoelison, M., & de Neys, W. (2020). "you're wrong!": The impact of accuracy feedback on the bat-and-ball problem. *Acta Psychologica*, *206*, 103042. https://doi.org/10.1016/j.actpsy.2020.103042

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus; Giroux.

Klauer, K. C., & Singmann, H. (2013). Does logic feel good? testing for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *39*, 1265–1273. https://doi.org/10.1037/a0030530

Kokis, J. V., Macpherson, R., Toplak, M. E., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology*, *83*(1), 26–52. https://doi.org/10.1016/S0022-0965(02)00121-2

Lanoë, C., Lubin, A., Houdé, O., Borst, G., & De Neys, W. (2017). Grammatical attraction error detection in children and adolescents. *Cognitive Development*, *44*, 127–138. https://doi.org/10.1016/j.cogdev.2017.09.002

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

Markovits, H., & Barrouillet, P. (2004). Introduction: Why is understanding the development of reasoning important? *Thinking & Reasoning*, *10*(2), 113–121. https://doi.org/10.1080/13546780442000006

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

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & cognition*, *17*(1), 11–17. https://doi.org/10.3758/bf03199552

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. https://doi.org/10.1037/0096-3445.130.4.621

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological bulletin*, *132*(2), 297–326. https://doi.org/10.1037/0033-2909.132.2.297

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

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

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

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

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. [PMID: 25530822]. *Judgment & Decision Making*, *7*(3), 332–359.

Reyna, V. F., & Brainerd, C. J. (2011). Dual processes in decision making and developmental neuroscience: A fuzzy-trace model [Special Issue: Dual-Process Theories of Cognitive Development]. *Developmental Review*, *31*(2), 180–206. https://doi.org/10.1016/j.dr.2011.07.004

Reyna, V. F., & Ellis, S. C. (1994). Fuzzy-trace theory and framing effects in children's risky decision making. *Psychological Science*, *5*(5), 275–279. https://doi.org/10.1111/j.1467-9280.1994.tb00625.x

Reyna, V. F., Rahimi-Golkhandan, S., Garavito, D. M. N., & Helm, R. K. (2017). The fuzzy-trace dual-process model. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 82–99). Routledge.

Shiffrin, R., & Schneider, W. (1977). Controlled and automatic human information processing: Ii. perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*, 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Singmann, H., Klauer, K. C., & Kellen, D. (2014). Intuitive logic revisited: New data and a bayesian mixed model meta-analysis. *PLoS One*, *9*(4). https://doi.org/10.1371/journal.pone.0094223

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

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

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

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

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

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

Supplementary material

A. Material rating

For base-rate problems, two categories were presented, with the first category supposed to be cued by the description (intended stereotypical association), while the second category was not. Participants had to rate to which extent they agreed with the association between each category and the description (e.g., "a clown is funny" and "an accountant is funny"). Figure S1 (left panel) shows the ratings for individual items.

For syllogisms, participants were asked to rate the believability of each conclusion. Four items had conclusions that were intended to be believable, four others were intended to be unbelievable. Figure S1 (right panel) illustrates the ratings for believable (first category) and unbelievable (second category) items.

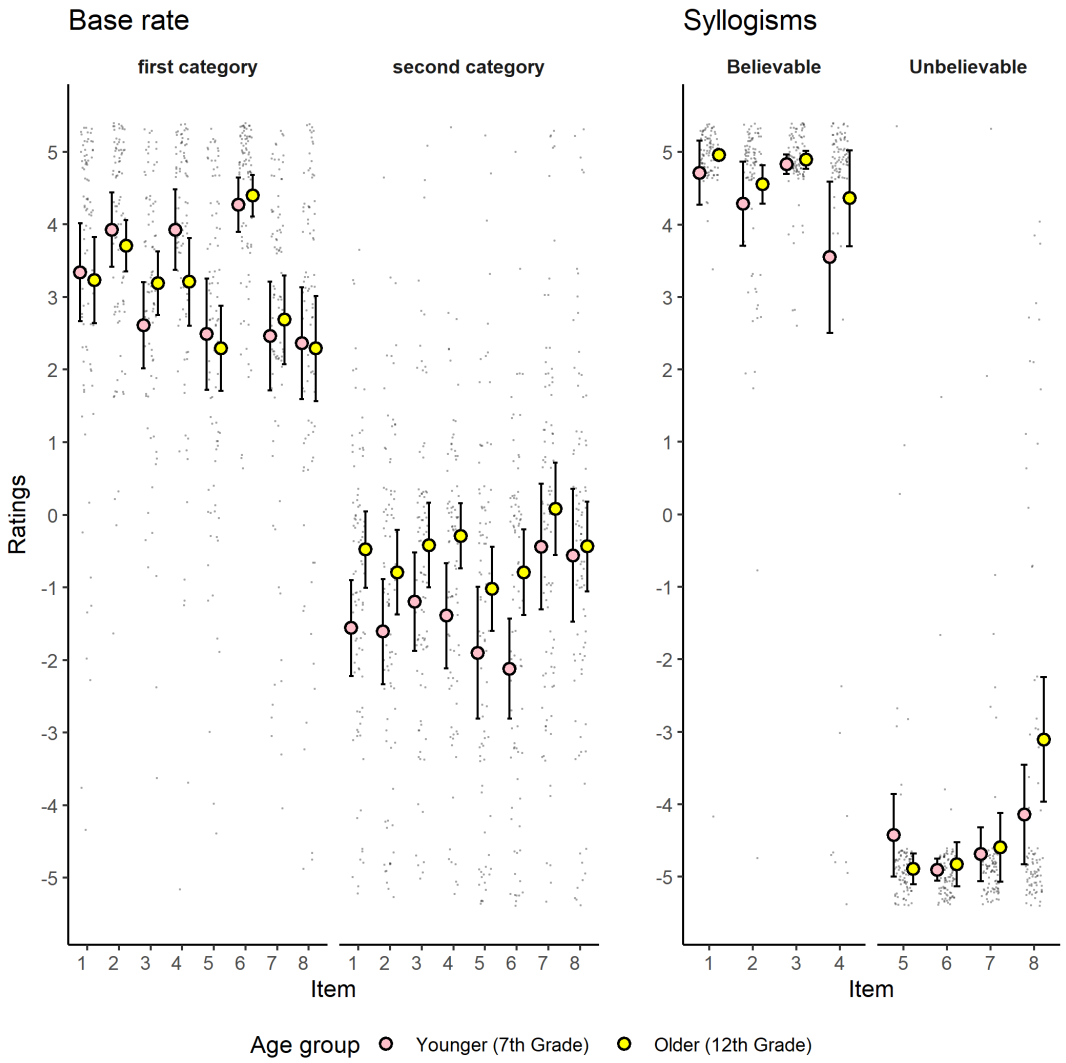

Figure S1. Individual ratings for base-rate problems and syllogisms. *Note.* Dots represent individual ratings. Circles and intervals represent average rating and 95% confidence interval.

B. Reasoning material (freely translated)

Base-rate problems

Conflict items:

1. *This study contains boxers and supermarket cashiers.*
   Person 'W' is strong.
   There are 5 boxers and 995 supermarket cashiers.
   Is Person 'W' more likely to be:
   -A boxer
   -A supermarket cashier?

2. *This study contains writers and construction workers.*
   Person 'C' is creative.
   There are 6 writers and 994 construction workers.
   Is Person 'C' more likely to be:
   -A writer
   -A construction worker?

3. *This study contains executive managers and humorists.*
   Person 'F' is robust.
   There are 996 boxers and 4 supermarket cashiers.
   Is Person 'F' more likely to be:
   -A boxer
   -A supermarket cashier?

4. *This study contains executive managers and humorists.*
   Person 'K' is funny.
   There are 997 executive managers and 3 humorists.
   Is Person 'K' more likely to be:
   -An executive manager
   -A humorist?

No-conflict items:

5. *This study contains flight attendants and prison officers.*
   Person 'M' is charming.
   There are 997 flight attendants and 3 prison officers.
   Is Person 'M' more likely to be:
   -A flight attendant
   -A prison officer?

6. *This study contains firemen and wealthy heirs.*
   Person 'D' is courageous.
   There are 996 firemen and 4 wealthy heirs.
   Is Person 'L' more likely to be:
   -A fireman
   -A wealthy heir?

7. *This study businessmen and garbage collectors.*
   Person 'D' is ambitious.
   There are 994 businessmen and 6 garbage collectors.
   Is Person 'D' more likely to be:
   -A businessman
   -A garbage collector?

8. *This study contains CEOs and gardeners.*
   Person 'S' is bossy.
   There are 995 CEOs and 5 gardeners.
   Is Person 'S' more likely to be:
   -A CEO
   -A gardener?

Neutral items:

1. *This study contains saxophonists and trumpetists.*
   Person 'F' is musical.
   There are 995 saxophonists and 5 trumpetists.
   Is Person 'F' more likely to be:

-A saxophonist
-A trumpetist?

2. *This study contains grandfathers and grandmothers.*
   Person 'G' is old.
   There are 5 grandfathers and 995 grandmothers.
   Is Person 'G' more likely to be:
   -A grandfather
   -A grandmother?

Syllogisms

1. *All flowers need water.*
   Roses need water.
   Roses are flowers.
   Does the conclusion follow logically? (BELIEVABLE, INVALID)

2. *Everything with a motor need oil.*
   Cars need oil.
   Cars have a motor.
   Does the conclusion follow logically? (BELIEVABLE, INVALID)

3. *All birds have wings.*
   Crows are birds.
   Crows have wings.
   Does the conclusion follow logically? (BELIEVABLE, VALID)

4. *Everything that is smokable is noxious.*
   Cigarettes are smokable.
   Cigarettes are noxious.
   Does the conclusion follow logically? (BELIEVABLE, VALID)

5. *All African countries are warm countries.*
   Spain is a warm country.

Spain is an African country.
Does the conclusion follow logically? (UNBELIEVABLE, INVALID)

6. *All meat-based products are edible.*
   Apples are edible.
   Apples are meat-based products.
   Does the conclusion follow logically? (UNBELIEVABLE, INVALID)

7. *All mammals can walk.*
   Whales are mammals.
   Whales can walk.
   Does the conclusion follow logically? (UNBELIEVABLE, VALID)

8. *All vehicles have wheels.*
   A boat is a vehicle.
   A boat has wheels.
   Does the conclusion follow logically? (UNBELIEVABLE, VALID)

NEUTRAL ITEMS

1. *All F are H.*
   All Y are F.
   All Y are H.
   Does the conclusion follow logically? (VALID)

2. *All L are P.*
   All Z are P.
   All Z are L.
   Does the conclusion follow logically? (INVALID)

C. Direction of change analysis: no-conflict items

As figure S2 shows, in both age groups, no-conflict trials were predominantly *11* for both tasks, with an age-related increase. Directional t-tests showed that the *11* increase for base-rate problems (young: $M = 81.5\%$, $SD = 26.4\%$; old: $M = 92.5\%$, $SD = 16.3\%$;

9% more) was significant, $t(126.95) = 3.27$, $p < .001$ (one-tailed), $r = .28$, as well as the *11* increase for syllogisms (young: $M = 82.7\%$, $SD = 24.3\%$; old: $M = 93\%$, $SD = 13.8\%$; 11.1% more), $t(120.30) = 3.35$, $p < .001$ (one-tailed), $r = .29$. These results parallel the accuracy findings and indicate that although there was a slight age-related increase, both age groups had little difficulty in giving intuitive correct responses when faced with no-conflict problems.

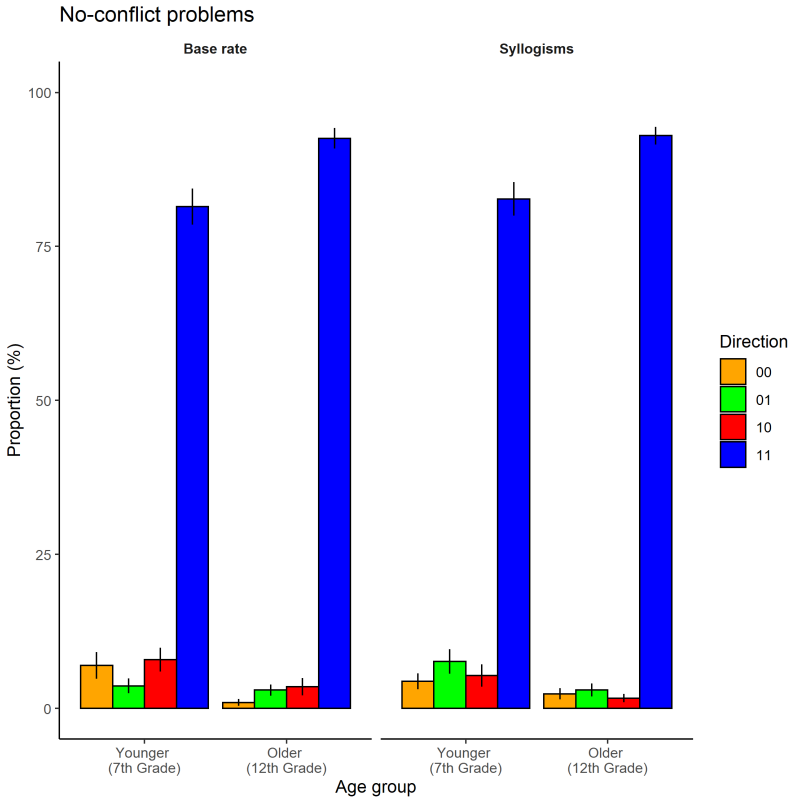

Figure S2. Proportions of direction of change categories for no-conflict problems. *Note.* Error bars represent standard error. Proportions for each direction were computed for each individual. Averages were then computed separately for each direction. *Note. 00* = both initial and final responses incorrect, *01* = incorrect initial response but correct final response, *10* = correct initial response but incorrect final response, and *11* = both initial and final responses correct.

D. Additional figures

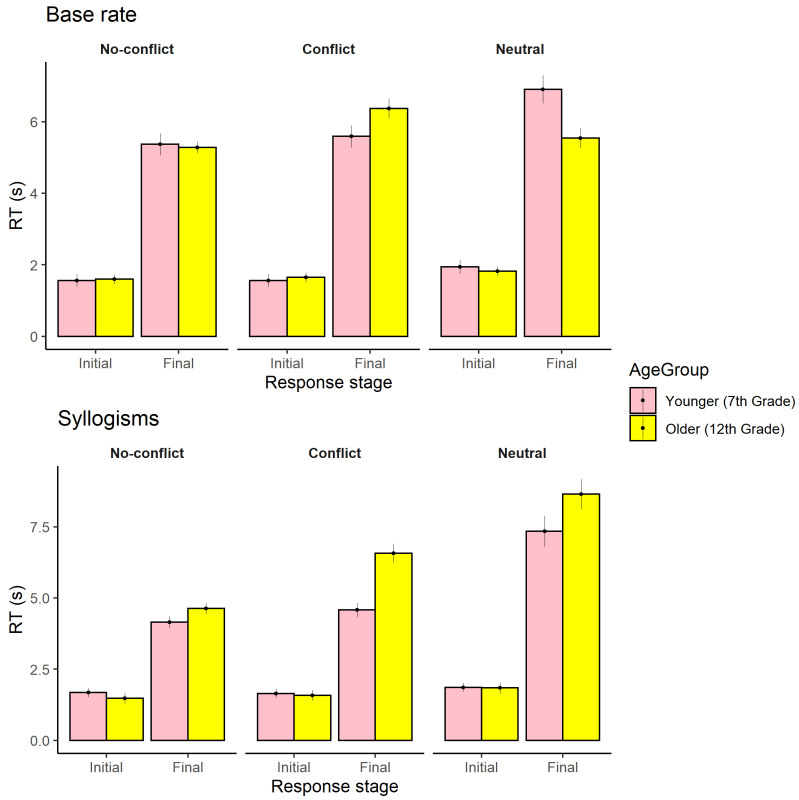

Figure S3. Response latencies for base-rate problems and syllogisms. *Note.* Error bars represent standard error. Individual log-transformed reaction times (RT) were averaged then back-transformed for ease of interpretation. For the final response stage, RT in each age group x problem type group, that deviated from the average by three times the *SD* or more, were excluded.

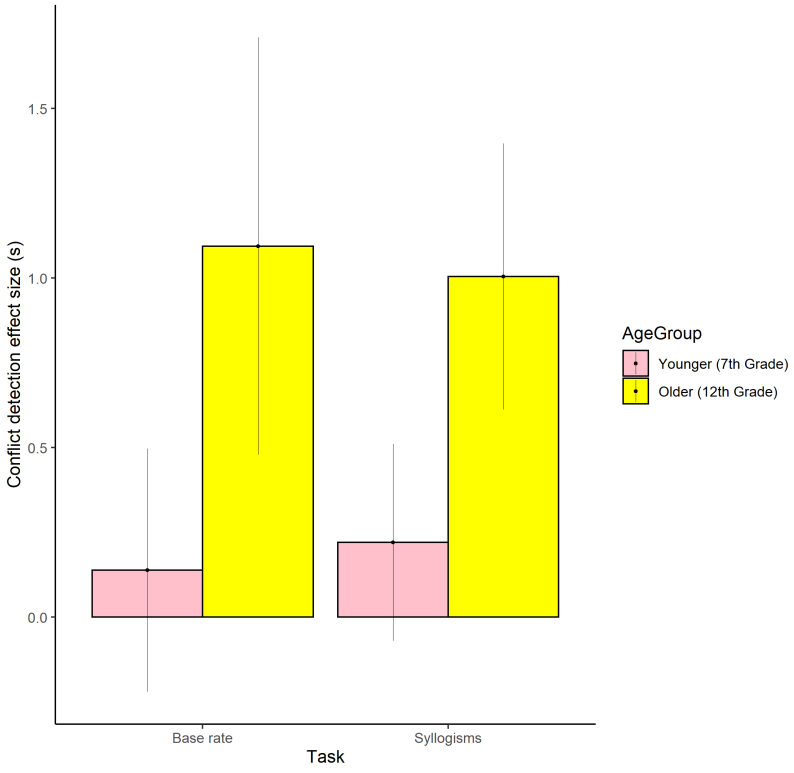

Figure S4. Conflict detection effect size. *Note.* Error bars represent standard error. Individual log-transformed reaction times (RT) were averaged then back-transformed to compute individual effect size (i.e., average RT incorrect conflict – average RT correct no-conflict). Only participants with at least one incorrect conflict item and one correct no-conflict item were included (i.e., $n = 103$ for the base-rate task and $n = 150$ for syllogisms). RT in each age group x problem type group, which were separated from the average by three times the *SD* or more, were excluded.