# Emergence of the Smart Intuitor:

# How Cognitive Ability Shapes Adolescent Reasoning

Laura Charbit[a,*], Esther Boissin[b], Matthieu Raoelison[a], Wim De Neys[a]

[a] Université Paris Cité, CNRS, LaPsyDÉ, F-75005 Paris, France

[b] Cornell University, Department of Psychology, Ithaca, NY, US 14853

[*] Corresponding author: laurajcharbit@gmail.com, 46 rue Saint-Jacques, 75005, Paris, France.

**Abstract**

From early adolescence, cognitive ability has been shown to predict reasoning performance. Recent studies indicate that adults with high cognitive ability tend to rely on accurate intuitions rather than deliberately correcting incorrect intuitions. However, it is unclear whether this pattern applies to adolescents. In the current study, we tested whether – as previously reported – younger adolescents are less likely to intuit correctly and whether cognitive ability is more predictive of correct intuitions or deliberate corrections in 7th graders and 12th graders. We used a two-response paradigm where participants gave a fast intuitive response under time pressure and cognitive load, followed by an unconstrained deliberate response. Results confirmed that younger adolescents produce fewer correct intuitions than older ones. Additionally, in younger adolescents, cognitive ability did not positively correlate with reasoning performance, while in older adolescents, it mainly predicted deliberation. The study suggests that the previously established association between sound intuiting and cognitive ability among adults emerges quite late in development.

**Keywords:** dual-process; logical intuitions; development; cognitive ability; two-response paradigm

## Introduction

Even though we may not always be aware of it, our judgment is often biased. Decades of research have shown that we frequently fail to apply basic logical, mathematical, and probabilistic principles (e.g., Kahneman, 2011). Instead of engaging in reflection, we tend to rely on intuitive impressions when reasoning. Sometimes, this intuitive thinking is useful, leading to a valid conclusion quickly and with little effort. At other times, however, it leads to answers that conflict with logical, mathematical, or probabilistic principles. Consider this problem, for example:

*"There are 995 dogs and 5 cats in an animal shelter. A pet is selected at random. This pet is described as very independent and fond of fish. Do you think the randomly selected pet is more likely to be a dog or a cat?"*

Intuitively, many people tend to conclude that the randomly chosen pet is more likely to be a cat based on stereotypical beliefs cued by the description. This could be a fair guess if the description were the only information available. However, you are also told that there are almost 200 times more dogs than cats. Since there are far more dogs at the animal shelter, and dogs can also be independent and fond of fish, the correct answer is that the randomly selected pet is more likely to be a dog. However, reasoners are often led astray by their intuition and fail to solve the problem (e.g., Kahneman & Tversky, 1973).

Traditionally, biased reasoning has been explained by the classic dual-process theory, which conceptualizes thinking as an interaction between intuitive and deliberative processing, often referred to as System 1 and System 2. Intuitive System 1 operates quickly and effortlessly, providing immediate responses. In contrast, deliberative System 2 is slower and more effortful, placing a burden on our cognitive resources (e.g., see Evans, 2008, for a review). In this framework, applying logical principles would require demanding deliberation (e.g., Evans, 2008; Kahneman, 2011). However, because people tend to minimize demanding computations, they stick to the intuitive incorrect response that quickly comes to mind (Kahneman, 2011). The few reasoners who manage to answer correctly (i.e., in line with logical principles) are expected to engage in deliberation to correct their erroneous intuitions (Evans, 2008; Kahneman, 2011; Sloman, 1996).

However, recent studies have challenged this traditional view by showing that correct answers can be intuitive and do not necessarily require deliberative correction (e.g., Bago & De Neys, 2017, 2019; Newman et al., 2017; Thompson et al., 2011). These studies use a two-response paradigm (Thompson et al., 2011) in which participants respond to the same reasoning problem twice. First, they have to answer as fast as possible with the first intuitive

response that comes to mind. Next, they can take as much time as they want to reflect on the problem and give a final response. To ensure that the initial response is generated intuitively, it often has to be generated under concurrent secondary task load and/or time pressure (Bago & De Neys, 2017; Newman et al., 2017). The key finding of these studies is that reasoners who produce a correct final response often already gave a correct intuitive response to begin with (Bago & De Neys, 2017, 2019; Newman et al., 2017; Raoelison et al., 2020; Thompson et al., 2011). Thus, contrary to the traditional assumption, more recent dual-process models suggest that sound reasoners can produce logical intuitions and do not necessarily require corrective deliberation to arrive at the correct answer (e.g., De Neys, 2012, 2023; De Neys & Pennycook, 2019). Given that correct responses can arise without deliberation, this raises important questions about their origins: Where do the intuitions come from (e.g., see commentaries on De Neys, 2023)?

Intuitions can be understood as highly specialized cognitive procedures that have been practiced to the point of automaticity and are executed autonomously when triggered by specific stimuli (e.g., Bago & De Neys, 2020; De Neys, 2012; Evans, 2019; Stanovich, 2018). From this perspective, logical intuitions develop through repeated exposure to key logico-mathematical principles, particularly throughout the school curriculum, where they are taught and practiced by young reasoners (De Neys, 2012). For example, in France, adolescents begin learning to estimate and manipulate probabilities in 7th grade and continue exercising these skills until the end of secondary school. As a result, sound adult reasoners may have automatized and instantiated the critical logical knowledge structures or "mindware" to such a degree that they can apply them without deliberation (Stanovich, 2018). Recent research supports this automatization hypothesis by showing that older adolescents (i.e., 12th graders) who have been more exposed to and practiced logico-mathematical principles than younger adolescents (7th graders), not only engage in more corrective deliberations but also demonstrate more accurate intuitions in probabilistic and syllogistic reasoning tasks (Raoelison et al., 2021).

Interestingly, correct intuitive responding in adults is associated with higher cognitive ability. While classical dual-process theory traditionally suggests that individuals with higher cognitive ability are more likely to respond correctly through deliberation (e.g., De Neys, 2006; De Neys & Verschueren, 2006; Evans & Stanovich, 2013; Kahneman, 2011; Stanovich, 2011; Stanovich & West, 1999, 2000; Toplak et al., 2011), recent research using the two-response paradigm has challenged this 'smart deliberator' perspective (e.g., Raoelison et al., 2020; Thompson et al., 2018). For example, Raoelison et al. (2020) have found that although cognitive ability was associated with the tendency to engage in deliberation and correct incorrect intuitions (i.e., changing from an incorrect initial to a correct final response), it primarily predicted having sound intuitions (i.e., correct final responses that were also preceded by

correct intuitive ones). Thus, adults with higher cognitive ability appear to be 'smart intuitors': they generate more accurate intuitions rather than relying solely on deliberation.

In adolescence, the relationship between cognitive ability and correct reasoning has also been documented across several reasoning tasks, including probabilistic reasoning, syllogistic reasoning, and resistance to framing (Toplak & Flora, 2020; Kokis et al., 2002; Toplak et al., 2014). These studies consistently report a moderate to strong correlation between cognitive ability and reasoning accuracy. However, unlike in adults, the nature of this association remains unclear. While cognitive ability in adults is known to primarily predict accurate intuitions (Raoelison et al., 2020), research on adolescents has not examined whether cognitive ability predicts intuitive reasoning, deliberative reasoning, or both. Existing research focuses on overall reasoning performance without disentangling these processes. This leaves open the question of whether cognitive ability plays different roles in reasoning development across developmental stages. Given that logical intuitions are still developing in younger individuals, cognitive ability may be more strongly associated with accurate deliberation rather than accurate intuition during adolescence.

Hence, our primary interest in the present paper was to investigate the relationship between cognitive ability (as measured here with a fluid intelligence test) and both intuitive and deliberative reasoning across development and to understand whether it varies between younger and older adolescents. To address this issue, we tested two groups of adolescents: younger (7th graders, age 12-13) and older adolescents (12th graders, age 17-18), at the beginning and end of secondary school, respectively. We assessed reasoning performance using the two-response paradigm in order to differentiate between intuitive and deliberative responses. We selected 7th graders as the younger group because previous studies have shown that they are capable of completing various reasoning tasks using a two-response paradigm (Raoelison et al., 2021; Boissin et al., in prep). We selected 12th graders because they are near the end of secondary education, where the hypothesized automatization process should be advanced and potentially approach adult levels. Critically, both age groups could be tested in a similar school context.

A secondary objective was to test the previously reported lower prevalence of sound intuition among younger adolescents (Raoelison et al., 2021). This finding has important potential implications since it suggests that developmental improvements in reasoning accuracy are at least partially driven by an improvement in the accuracy of our intuitions. However, to the best of our knowledge, only one study (Raoelison et al., 2021) has contrasted the prevalence of sound intuition between younger and older adolescents. The present study allowed us to test the robustness of these findings.

To test the generalizability of our results, participants solved two classic bias tasks reflecting distinct heuristic and logico-mathematical principles: the base-rate task (Kahneman

& Tversky, 1973), and the bat-and-ball task ( "*A bat and a ball cost together $1.10. The bat costs $1 more than the ball. How much does the ball cost?*" Frederick, 2005). In the latter item, the problem cues a heuristic response that conflicts with the correct answer based on mathematical rules while the correct solution to the base-rate relies on the use of probabilistic principles.

## Method

### Preregistration and data availability

The study design and research question were preregistered on the AsPredicted website (https://aspredicted.org) and stored on the Open Science Framework (https://osf.io/5jnsc/) where all data and material can be accessed. No specific analyses were preregistered.

### Deviation from preregistration

We initially preregistered two reasoning tasks: the base-rate task and the bat-and-ball task. However, performance on the bat-and-ball task was at floor level across both age groups, with only 10 younger adolescents and 21 older adolescents providing at least one correct response (either initial or final). As a result, internal consistency (KR-20) could not be computed in this age group due to the lack of response variability. Given these limitations, and following reviewer recommendations, we focus our main analyses on the base-rate task and present the bat-and-ball task as exploratory data in the Supplementary Materials.

### Participants

We recruited 324 students from French secondary schools, divided into two age groups: 163 younger adolescents in 7th grade (*Mean age* = 12.31, *SD* = 0.63; 82 female, 4 prefer not to say) and 161 older adolescents in 12th grade (*Mean age* = 17.47, *SD* = 0.66; 82 female, 6 other, 3 prefer not to say). Consent was obtained from all students, as well as from a legal guardian of those under the age of 18.

We determined our sample size using an a priori power analysis. In adults, the correlations between cognitive ability and correction tendency ($r$ = .22) and between cognitive ability and correct intuitive responding ($r$ = .44) fall within the medium to medium-strong range (Raoelison et al., 2020). Based on this, we aimed to detect medium correlations in our study. According to G*Power 3 (Faul et al., 2007), a sample size of 159 participants per age group is required to achieve a power of 0.80 in detecting medium correlations ($r$ = .22 and $\alpha$ = .05). We recruited a slightly larger number of participants than initially determined by the power analysis to account for potential data loss.

## Material

### *Reasoning task*

**Base-Rate items.** We selected eight base-rate problems that were originally adapted in French by Raoelison et al. (2021) from the work of Pennycook et al. (2015). Each problem provided a description of the composition of a sample (e.g., "*This study contains clowns and accountants*"), base-rate information (e.g., "*There are 995 accountants and 5 clowns."*) and a description designed to cue a stereotypical association (e.g. "*Person 'L' is funny*.") Participants were asked to indicate to which group the person most likely belonged.

The problem presentation followed that of Pennycook et al. (2015) with descriptive information and base-rates presented serially, and the amount of text minimized. First, participants received the names of the two groups in the sample (e.g., "*This study contains clowns and accountants*"). The descriptive information (e.g., "*Person 'L' is funny.*") was then displayed beneath the first sentence, which remained on the screen. The descriptive information specified a neutral name (e.g.,"*Person L"*) and a single-word personality trait (e.g., "*funny*") intended to trigger the stereotypical association. Finally, participants received the base-rate probabilities (e.g., "*There are 995 accountants and 5 clowns."*). The following illustrates the full problem format:

*This study contains clowns and accountants.*
*Person 'L' is funny.*
*There are 995 accountants and 5 clowns.*
*Is Person 'L' more likely to be:*

- *A clown*
- *An accountant*

Pennycook et al. (2015) pre-tested the material to ensure that the selected words consistently triggered the intended stereotypical associations without being overly diagnostic. As Bago and De Neys (2017) clarified, the importance of such a non-extreme and moderate association is not trivial. Note that we label the response that is in line with the base-rates as the correct response. Critics of the base-rate task (e.g., Gigerenzer et al., 1988; Barbey & Sloman, 2007) have long pointed out that if reasoners adopt a Bayesian approach and combine the base-rate probabilities with the stereotypical description, this can lead to interpretative complications when the description is extremely diagnostic. For example, imagine that we have an item with males and females as the two groups and give the description that Person 'A' is 'pregnant'. Now, in this case, one would always need to conclude that Person 'A' is a woman, regardless of the base-rates. The more moderate descriptions

(such as ‘kind’ or ‘funny’) help to avoid this potential problem. In addition, the extreme base-rates (i.e., 997/3, 996/4, 995/5) that were used in the current study further help to guarantee that even a very approximate Bayesian reasoner would need to pick the response cued by the base-rates (see De Neys, 2014). Raoelison et al. (2021) also pre-tested the translated material to ensure that it cued the intended stereotypes among both younger and older French adolescents. Their results showed that the material worked as intended and was equally familiar and diagnostic across both age groups.

Half of the problems were featured in their standard "conflict" version and the other half in their no-conflict version. In conflict items the base-rate probabilities and the stereotypical information cued opposite responses, while in no-conflict items they cued the same response (i.e., the description triggered a stereotypical trait of a member of the largest group). No-conflict items served as control problems to ensure that participants were focused on the task and not guessing randomly. For instance:

*This study contains clowns and accountants.*
*Person ’L’ is funny.*
*There are 5 accountants and 995 clowns.*
*Is Person ’L’ more likely to be:*

- *A clown*
- *An accountant*

Two sets of problems were used to counterbalance problem content. By switching the base-rates of the two groups, the conflict problems in one set became the no-conflict problems in the other, and vice-versa. As a result, set A contained four conflict and four no-conflict items, while set B contained the matching four no-conflict and four conflict items. Participants were randomly assigned to one of the two sets. The presentation order of the problems was randomized for each participant. In total, each participant solved four conflict and four no-conflict items.

**Two-response format.** Participants responded to each base-rate problem using a two-response paradigm, where they first provided a ‘fast’ answer, directly followed by a second ‘slow’ answer (Thompson et al., 2011). This method allowed us to capture both an initial "intuitive" response and then a final "deliberate" one. To minimize the possibility that deliberation was involved in producing the initial ‘fast’ response, participants had to provide their initial answer within a strict time limit while performing a concurrent cognitive load task (see Bago & De Neys, 2017, 2019). The load task was based on the dot memorization task (Miyake et al., 2001) given that it had been successfully used to burden executive resources

during reasoning tasks (e.g., De Neys, 2006; Franssens & De 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 their initial (intuitive) response to the problem, participants were shown four different patterns and had to identify the one that they had memorized (see Bago & De Neys, 2019, for more details).

The initial response deadline was set at 3 seconds, based on pretesting by Bago and De Neys (2017). The allotted time corresponded to the time required to read the problem, the question and answer alternatives, move the mouse, and select an answer among the possibilities (see Bago & De Neys, 2017, for details). Furthermore, participants were also under a secondary task load when giving their initial response, making intuitive responding even more challenging. Obviously, the time limit and cognitive load were applied only for the initial response, and not for the final one where participants were allowed to deliberate (see below).

**Two-response format and development.** The two-response paradigm can be quite challenging, especially for younger adolescents, who may have difficulty reading and comprehending the problems under strict time constraints and cognitive load. However, several studies (Raoelison et al., 2021; Boissin et al., in prep) have shown that we can use the two-response procedure with young and old adolescents on various reasoning tasks. In these studies, the same deadline was used for both younger (around 12 years old) and older adolescents (around 17 years old) when contrasting the intuitive reasoning performance of these age groups. Although both the deadline and the load memorization task were challenging, (younger) participants were able to meet them on the vast majority of trials. Similarly, in the current study, (younger) participants met constraints on most trials (see Trial exclusions below). Moreover, no-conflict accuracies for initial responses were very high among younger adolescents ($M$ = 90.42%, $SD$ = 9.85%), indicating that 7th graders were able to solve base-rate problems under these stringent conditions and refrained from mere random guessing. These results suggest that the two-response procedure can be effectively used with the base-rate task in a young adolescent sample.

### *Cognitive ability task*

Raven's Standard Progressive Matrices task has been widely used to measure “general cognitive ability” of different age groups (SPM; Raven et al., 1998a). A Raven problem presents a 3 × 3 matrix of complex visual patterns with a missing element, requiring one to choose the only pattern matching both row- and column-wise from six or eight alternatives. We used the 15-item short form of the SPM developed by Langener et al. (2021) for adolescents. To address a potential ceiling effect for the older age group, we added a more challenging item from Raven’s Advanced Progressive Matrices (APM; Raven et al., 1998b) as suggested by

Bilker et al. (2012). Each participant's cognitive ability score was defined by their total accuracy across the 16 matrices.

**Procedure**

The experiment was run individually on iPads. The participants were tested in their classrooms in groups under the supervision of at least one teacher and one experimenter. At the start, participants were informed that the study would take about 30 minutes and consist of three tasks. Short transition messages (e.g., "*Well done, you have finished part 1. Please click on Next when you are ready to start part 2.*") marked progress between tasks.

The two reasoning tasks (see Supplementary Materials, Section A, for details on the bat-and-ball task) were presented first, in a randomized order. Before beginning the first reasoning task, participants read detailed instructions on the two-response format (see Raoelison et al., 2021). For the base-rate task, they also received the following explanation:

*"In a big research project a large number of studies were carried out where short personality descriptions of the participants were made. In every study there were participants from two population groups (e.g., carpenters and policemen).*
*In each study one participant was drawn at random from the sample. You'll get to see a personality trait for this randomly chosen participant. You'll also get information about the composition of the population groups tested in the study in question.*
*You'll be asked to indicate to which population group the participant most likely belongs."*

To familiarize themselves with the initial response deadline and the two-response procedure, participants first solved two no-conflict practice reasoning problems. Next, they solved two practice cross memorization items (without concurrent reasoning problem). Finally, they revisited the two earlier practice reasoning problems under cognitive load.

As illustrated in Figure 1, base-rate trials began with a fixation cross displayed for 2000 ms, followed by a description of the sample population (e.g., "*This study contains clowns and accountants.*") for another 2000 ms. Next, the target pattern for the memorization task appeared for 2000 ms. Afterwards, the first two sentences of the problem, which included the descriptive adjective (e.g. "*Person 'L' is funny.*") were displayed for 2000 ms. Finally, the full problem, including the base-rate information (e.g., "*There are 995 accountants and 5 clowns.*") and the answer options, were presented. At this point, participants had 3000 ms to select their initial answer from two choices (see Figure 1). After 2000 ms, the background of the screen turned yellow to signal that time was running out.

If participants had not provided an answer before the time limit, they were given a reminder that it was important to provide an answer within the time limit on subsequent trials. Then, they were presented with four visual matrices and had to choose the one that they had previously memorized. They received feedback as to whether their memory-response was correct. If the answer was not correct, they were reminded that it was important to perform well on the memory task on subsequent trials. Finally, the same reasoning problem was presented again, and participants were asked to provide a final deliberate answer (with no time limit). The color of the answer options was green during the initial 'intuitive' response, and blue during the final 'deliberate' 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 were also colored as the answer options.

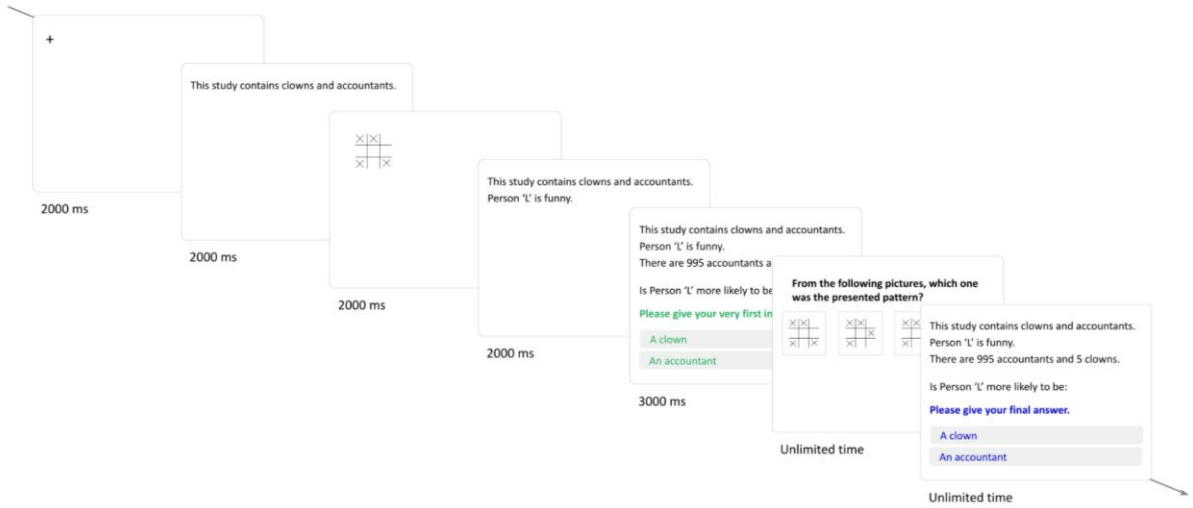


**Figure 1.** Time course of a complete two-response trial for the base-rate task.

Finally, after the two reasoning tasks, participants were invited to complete the cognitive ability task. Participants were first shown a Raven item along with the solution and basic explanations, followed by a practice item for which they received feedback. The instructions were as follows:

*Please read these instructions carefully!*
*Here is a picture with a missing fragment.*
*Your task is to pick the correct fragment to complete the picture.*
*You have to pick the fragment that matches both horizontally AND vertically.*
*You'll start with a practice problem. There is no time limit. Once you have picked your answer, you can select it.*
*Please click on Next to start practice.*

The 16 Raven items were then presented serially, one after the other. Participants had to provide an answer to each item. There was no time limit for providing an answer.

At the very end of the experiment, participants were shown the standard bat-and-ball problem and were asked whether they had seen it before. They were also asked to enter the solution. Finally, participants completed a page of demographic questions.

## Results

The data were processed and analyzed using the R software (R Core Team, 2024) and the following packages (in alphabetical order): cocor (Diedenhofen & Musch, 2015), dplyr (Wickham et al., 2023),ez (Lawrence, 2016), ggplot2 (Wickham, et al., 2024), ggpubr (Kassambara, 2023a), psych (Revelle, 2025), rstatix (Kassambara, 2023b), and tidyr (Wickham et al., 2024).

### Trial exclusion

We discarded trials in which participants either failed to respond before the deadline or did not successfully complete the load memorization task, as we could not guarantee that their initial responses on these trials were free from deliberation (Bago & De Neys, 2017). Table 1 details the excluded trials by age group. Ultimately, we analyzed 82.30% of base-rate trials for older adolescents and 78.26% for younger adolescents. 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).

**Table 1.** Percentage of excluded and analyzed trials for the base-rate task by age group.

| Age group | Excluded trials | | Analyzed trials (%) |
|---|---|---|---|
| | Missed deadline (%) | Failed load for the remaining trials (%) | |
| Younger adolescents | 5.37 | 16.37 | 78.26 |
| Older adolescents | 3.11 | 14.60 | 82.30 |

**Note.** Excluded trials were calculated in two steps. First, the trials with a missed deadline were discarded. Next, from the remaining trials, those with a failed load were excluded.

We also measured the average individual contribution. For the base-rate task, on average older adolescents contributed 3.17 (*SD* = 0.96) conflict trials out of four and 3.25 (*SD* = 0.87) no-conflict trials out of four; younger adolescents contributed 2.93 (*SD* = 1.05) conflict trials and 3.03 (*SD* = 0.99) no-conflict trials.

## Reasoning performance

### *Accuracy*

The base-rate task showed good internal consistency for final responses in both older adolescents (set 1: KR-20 = .91; set 2: KR-20 = .87) and younger adolescents (set 1: KR-20 = .85; set 2: KR-20 = .87). However, the introduction of a strict deadline and cognitive load during the initial response phase led to lower reliability, with a slight decrease in older adolescents (set 1: KR-20 = .84; set 2: KR-20 = .79) and a more pronounced decrease in younger adolescents (set 1: KR-20 = .65; set 2: KR-20 = .61).

Base-rate conflict accuracies ranged from 0% to 100% in both age groups, exhibiting a bimodal distribution with peaks at 0% and 100% for both initial and final responses. This pattern (see Supplementary Materials, Section B, for distribution plots) is consistent with previous findings on the base-rate task (e.g., Pennycook et al., 2022). Figure 2A shows the performance for both younger and older adolescents for both initial intuitive and final deliberative responses. Visual inspection indicates that older adolescents outperformed younger adolescents on conflict items at both the initial and final response stages. In addition, older adolescents improved from the initial (*M* = 26.72%, *SD* = *15.64*%) to the final response stage (*M* = 40.10%, *SD* = *14.53*%), whereas younger adolescents showed little improvement after deliberation (initial: *M* = 22.03%, *SD* = 19.61%; final: *M* = 25.52%, *SD* = 13.22%). We ran a 2 (age group: younger or older) × 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. The analysis revealed a significant main effect of age group, $F(1, 318) = 6.13$, $p = .013$, $\eta^2_G = .016$, a significant main effect of response stage, $F(1, 318) = 23.53$, $p < .001$, $\eta^2_G = .012$, and a significant interaction between age group and response stage, $F(1, 318) = 8.09$, $p = .005$, $\eta^2_G = .004$. Furthermore, post hoc tests with Bonferroni correction indicated that the improvement between the initial and the final response stages was significant for older adolescents, $t(159) = 5.31$, $p < .001$, $d = 0.42$, but not for their younger counterparts, $t(159) = -1.46$, $p = .148$, $d = -0.12$. These findings indicate that older adolescents improve their performance from the initial to the final response stage when given the opportunity to deliberate, whereas younger adolescents do not yet demonstrate this benefit from deliberation.

Regarding the no-conflict performance, Figure 2A also shows that, accuracy was high for both older adolescents (initial: *M* = 94.71%, *SD* = 7.17%; final: *M* = 95.02%, *SD* = 6.90%) and younger adolescents (initial: *M* = 90.42%, *SD* = 9.85%; final: *M* = 91.41%, *SD* = 8.23%). T-tests confirmed that the performance of both age groups was significantly above chance (e.g., 50%) at both the initial and final response stages, all $p < .001$. This indicates that both younger and older participants were able to read and process the material and did not respond randomly on the base-rate task.

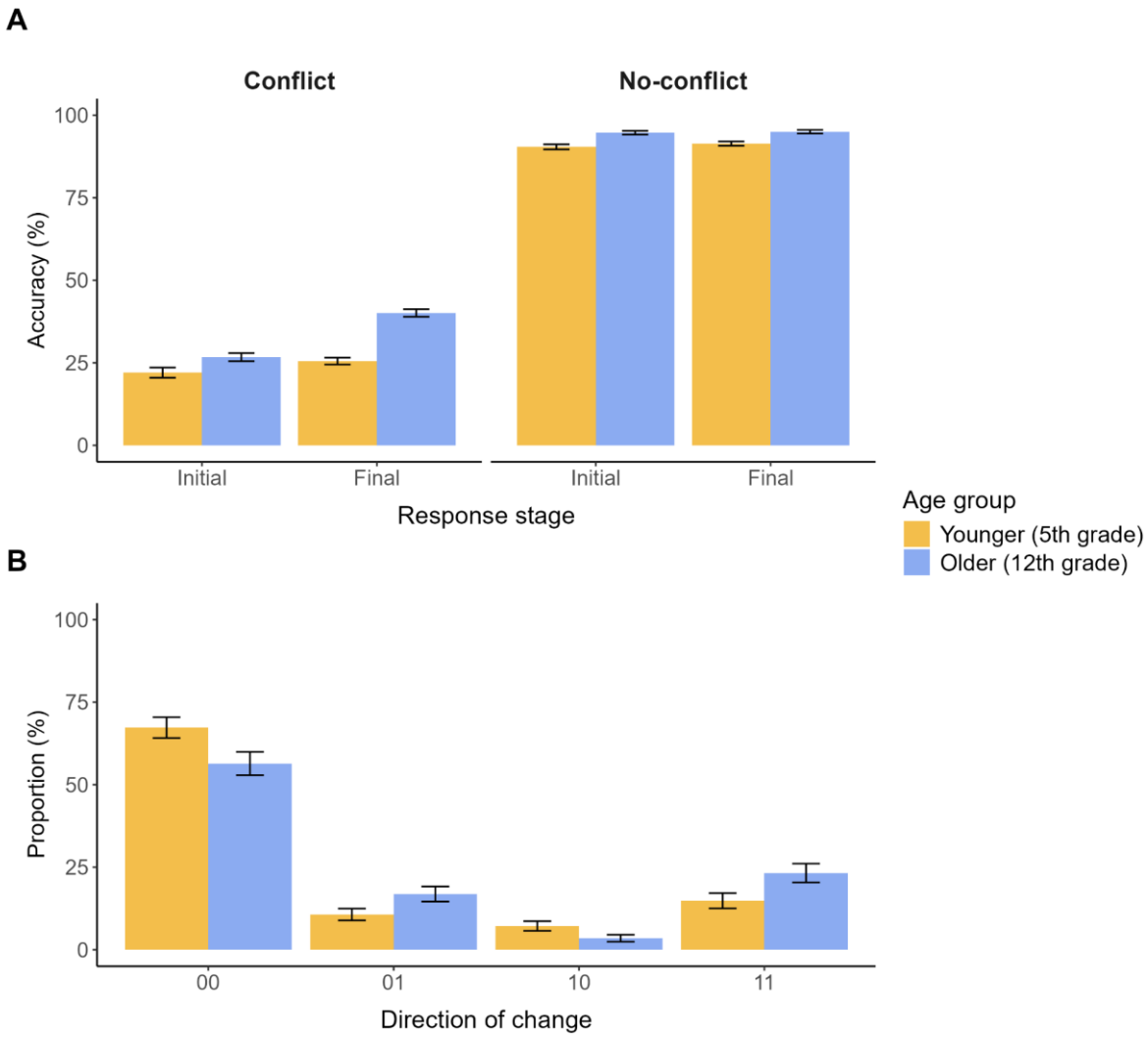


**Figure 2.** Accuracy and Proportions of direction of change for the base-rate task. Error bars represent Standard Error of the Mean (SEM). (A) Mean response accuracy for conflict items and no-conflict items as a function of response stage. (B) Proportions of direction of change (i.e., 00 = both initial 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) categories for conflict items.

***Direction of change***

To better understand how people changed (or did not change) their responses after deliberation, we performed a direction of change analysis for the conflict items (Bago & De Neys, 2017) for each age group. Specifically, each trial is composed of two responses, the initial 'intuitive' one (with time and load constraints) and the final 'deliberate' one. Correct responses are labeled '1' and incorrect responses are labeled '0'. Hence, each trial can result in one of these four patterns: "00" pattern, incorrect response at both response stages; "11" pattern, correct response at both response stages; "01" pattern, initial incorrect and final correct responses; "10" pattern, initial correct and final incorrect responses. Proportions for each direction were computed for each participant. Averages were then computed separately for each direction. Figure 2B shows the direction of change distribution for each age group (younger vs. older adolescents).

As Figure 2B shows, both older and younger adolescents primarily provided 00 response patterns (younger adolescents: $M$ = 67.29%, $SD$ = 39.80%; older adolescents: $M$ = 56.41%, $SD$ = 44.61%). However, older adolescents gave significantly fewer 00 response patterns than younger adolescents, $t(318) = 2.30$, $p = .022$, $d = .26$. They also provided significantly more 11 responses (younger adolescents: $M$ = 14.84%, $SD$ = 29.40%; older adolescents: $M$ = 23.23%, $SD$ = 36.07%), $t(318) = -2.28$, $p = .023$, $d = -.25$ as well as 01 responses (younger adolescents: $M$ = 10.68%, $SD$ = 22.57%; older adolescents: $M$ = 16.88%, $SD$ = 28.84%), $t(318) = -2.14$, $p = .033$, $d = -.24$, compared to younger adolescents. Consistent with previous research by Raoelison et al. (2021), our results indicate that older adolescents are more likely than their younger counterparts to provide correct responses on the base-rate task, both from intuition and after further deliberation.

To make sure that participants did not deliberate during the initial response stage, we excluded around 20% of trials in both age groups. In theory, this could have artificially boosted the proportion of 11 responses. That is, if these excluded trials would be specifically of the 01 or 00 type. 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 pointed to similar results for both younger (00: $M$ = 66.88%, $SD$ = 38.14%; 01: $M$ = 11.67%, $SD$ = 21.74%; 11: $M$ = 15.42%, $SD$ = 27.90%; 10: $M$ = 6.04%, $SD$ = 14.45%) and older adolescents (00: $M$ = 56.56%, $SD$ = 43.06%; 01: $M$ = 16.46%, $SD$ = 27.85%; 11: $M$ = 23.65%, SD = 35.58%; 10: $M$ = 3.33%, $SD$ = 10.83%).

### Cognitive ability performance

The Raven's Matrices demonstrated good internal consistency across age groups, with KR-20 values of .78 for younger adolescents and .81 for older adolescents. Scores ranged from 12.5% to 100% in both groups, with skewness and kurtosis values falling within acceptable limits (younger: skewness = 0.17, kurtosis = 2.31; older: skewness = -0.16, kurtosis = 2.32), indicating no substantial deviations from normality (see Supplementary Materials, Section B, for distribution plots). Older adolescents achieved an average accuracy of 61.37% (*SD* = 22.47%), significantly higher than the 52.45% (*SD* = *21.11*%) of younger participants, $t(322) = -3.68$, $p < .001$, $d = -.41$.

### Cognitive ability correlations

We correlated reasoning performance—initial and final accuracies, along with the direction of change category—with cognitive ability score[1].

To evaluate the potential influence of outliers on our correlation analyses, we computed Cook's distance for each observation, using a cutoff of 0.7 as recommended by McDonald (2002). No data points exceeded this threshold (all D < 0.7), indicating that our results were not driven by any single data point.

#### *Accuracy correlations*

Table 2 shows the correlations between cognitive ability and reasoning accuracy on the base-rate task. Older adolescents' cognitive ability correlated positively with final accuracy, $r(158) = .23$, $p = .004$, but not with initial accuracy, $r(158) = .09$, $p = .273$. A Steiger-Williams's t-test for dependent, overlapping correlations indicated that the two coefficients differed significantly, $t(157) = -2.31$, $p = .022$. This suggests that, in older adolescents, cognitive ability is more strongly associated with final (deliberate) accuracy than with initial (intuitive) accuracy.

In contrast, younger adolescents showed no significant positive correlation between cognitive ability and reasoning accuracy at either the final stage, $r(158) = .03$, $p = .712$, or the initial stage, $r(158) = -.14$, $p = .081$. A Steiger-Williams's t-test for dependent, overlapping correlations indicated that the two correlation coefficients differed significantly, $t(157) = -2.55$, $p = .012$. This pattern suggests that, for younger adolescents, higher cognitive ability is not associated with better reasoning performance at either response stage. If anything, younger adolescents tended to show a negative correlation trend for initial responses, suggesting that

[1] We also re-ran these correlations while including all missed load and missed deadline trials (by coding initial responses with a missed deadline as '0' in trials, and retaining the original initial response for trials with a failed load). The analysis yielded similar results for both younger and older adolescents.

those with lower cognitive ability were more likely to select correct responses, possibly reflecting increased guessing among participants lower in cognitive capacity.

To determine whether the link between cognitive ability and reasoning varies with age, we compared correlation coefficients across age groups using Fisher r-to-z tests for independent correlations. For initial responses, the correlation with cognitive ability differed significantly between younger and older adolescents, $z$ = -2.01, $p$ = .045. For final responses, the difference between age groups approaches significance, $z$ = -1.78, $p$ = .076. These results suggest that the association between cognitive ability and accuracy at the initial (intuitive) stage is relatively stronger in older than in younger adolescents, even though the within-group correlations are small and do not reach significance. This age difference becomes only marginal once participants have time to deliberate and produce a final answer.

**Table 2.** Correlations between cognitive ability and base-rate accuracy at the initial and final response stages for conflict items.

| Task | Age group | Response stage | *r* | *p* | df |
|---|---|---|---|---|---|
| Base-Rate | Younger | Initial | -.14 | .081 | 158 |
| | | Final | .03 | .712 | 158 |
| | Older | Initial | .09 | .273 | 158 |
| | | Final | .23** | .004 | 158 |

* $p$ < .05; ** $p$ < .01; *** $p$ < .001.

***Direction of change correlations***

To disentangle the relationship between cognitive ability and sound intuition from its relationship with corrective deliberation, we examined correlations between cognitive ability and the 01 (i.e., correct response only at the deliberate stage) and 11 (i.e. correct response from the intuitive stage) patterns. Table 3 presents these key correlations for the base-rate task. Older adolescent's cognitive ability correlated with 01 responses, $r$(158) = .22, $p$ = .005, but not with 11 responses, $r$(158) = .10, $p$ = .220. A Steiger-Williams's t-test for dependent, overlapping correlations showed that the difference between the two correlations was not significant, $t$(157) = 1.05, $p$ = .297. This suggests that while cognitive ability is significantly linked to corrective deliberation in older adolescents, this link is not significantly stronger than the one with sound intuition.

In contrast, younger adolescents showed no positive correlation between cognitive ability and 01 responses (i.e., correct response only at the deliberate stage), $r$(158) = .06, $p$ =

.478, nor between cognitive ability and 11 responses (i.e., correct response from the intuitive stage), $r(158) = -.01$, $p = .949$. A Steiger-Williams's t-test for dependent, overlapping correlations showed that the difference between the two correlations was not significant, $t(157) = 0.57$, $p = .572$.

To determine whether the link between cognitive ability and reasoning varies with age, we compared correlation coefficients across age groups using Fisher's r-to-z test for independent correlations. For 01 responses, the correlation with cognitive ability did not differ significantly between younger and older adolescents, $z = -1.47$, $p = .141$. Similarly, for 11 responses, the difference between age groups was also nonsignificant, $z = -0.91$, $p = .362$. These results indicate that although cognitive ability was associated with corrective deliberation in older adolescents, we did not detect a significantly stronger association than in younger adolescents. Likewise, the lack of correlation between cognitive ability and sound intuition was consistent across age groups.

**Table 3.** Correlations between cognitive ability and base-rate direction of change categories for conflict items.

| Task | Age group | Direction of change category | *r* | *p* | df |
|---|---|---|---|---|---|
| Base-Rate | Younger | 01 | .06 | .478 | 158 |
| | | 11 | -.01 | .949 | 158 |
| | Older | 01 | .22** | .005 | 158 |
| | | 11 | .10 | .220 | 158 |

* $p < .05$; ** $p < .01$; *** $p < .001$.

## Discussion

This study aimed to investigate how reasoning performance develops during adolescence and how it relates to cognitive ability. We contrasted the reasoning performance of younger (7th graders) and older adolescents (12th graders) on two classic reasoning tasks: the base-rate task and the bat-and-ball task. We used a two-response paradigm to differentiate correct intuition from corrective deliberation. Participants provided two consecutive responses: an initial "intuitive" response followed by a final "deliberative" one. Our goals were twofold: (1) to test the robustness of the developmental trend showing that older adolescents produce more correct intuitions than younger ones, and (2) to examine the link between cognitive ability and both intuitive and deliberative reasoning across adolescence.

First, regarding the developmental trajectory of correct intuitions, results showed that older adolescents produced more correct intuitions (i.e., “11” responses) than their younger counterparts. This replicates the findings of Raoelison et al. (2021) on the base-rate task. This robust increase in correct intuitive responding with age supports the automatization hypothesis. That is, it lends credit to the idea that logical intuitions result from the repeated exposure to key logico-mathematical principles, throughout the school curriculum (De Neys, 2012). With enough repetition, they might become so instantiated that they can be applied effortlessly, without the need for deliberation (Stanovich, 2018).

Second, we examined whether cognitive ability is more closely associated with intuitive or deliberative reasoning during adolescence. Regarding the younger adolescents, we found no indication whatsoever of a positive correlation between cognitive ability and reasoning—neither intuitive nor deliberative. The absence of a link with intuitive reasoning is not necessarily surprising, given that young adolescents are not yet expected to have developed logical intuitions due to limited exposure to these principles at the start of secondary education. However, the lack of relationship with deliberative reasoning is more unexpected, as previous studies have shown a positive association between cognitive ability and performance on heuristic and bias reasoning tasks in adolescents of similar age (Toplak & Flora, 2020; Toplak et al., 2014; Kokis et al., 2002). One possible explanation for this discrepancy is the low prevalence of correct responses among younger adolescents in our study, which may have limited the variability needed to detect a correlation. Additionally, the heuristic and bias tasks used in this study differ from those in previous research, possibly contributing to the divergent findings. Nevertheless, these results suggest that the positive association between cognitive ability and reasoning reported in prior studies may not be as consistent or generalizable in younger adolescents as previously assumed.

In older adolescents, cognitive ability predicted deliberative reasoning but not sound intuiting. The link with deliberative reasoning fits with numerous previous studies with late adolescents and adults (e.g., De Neys, 2006; Stanovich, 2011; Stanovich & West, 2000; Toplak et al., 2011). The absence of a link with intuitive reasoning indicates that, unlike findings with adults (Raoelison et al., 2020), there was no indication that cognitive ability was a better predictor of sound intuiting than deliberation. This difference may indicate that older adolescents’ application of logical principles is less instantiated than that of adults and that the automatization of this process continues into young adulthood. In other words, the “smart intuitor” pattern (Raoelison et al., 2020) where cognitive ability mainly predicts sound intuiting would only emerge fairly late in development. Until this point, accurate reasoning will often require effortful, deliberate correction of erroneous intuitions—and those higher in cognitive ability will be more likely to complete it successfully.

Notably, these correlations were only observed in the base-rate task. Given that not all logico-mathematical principles may be practiced to the same extent (or at the same time), potential differences between tasks and the specific logical principles they engage cannot be ruled out. To examine the generalizability of the relationship between cognitive ability and reasoning in adolescence, we included the bat-and-ball task. Unfortunately, the task proved too difficult, resulting in data with insufficient variability to draw meaningful conclusions. Nonetheless, this highlights the importance of incorporating a broader range of individual reasoning tasks in future studies.

While our overall sample size was sufficient to detect medium-sized correlations between cognitive ability and reasoning within each age group, it did not provide enough power to detect similar differences in the strength of these correlations. Notably, the correlation between cognitive ability and deliberative correction (i.e., “01” responses) was not significantly stronger in older adolescents than in younger ones. However, detecting such differences would require a much larger sample size, and given the practical constraints of in-person school testing, this was not feasible in the present study.

In sum, taken together, our findings generally support the idea that correct logical intuitions begin to emerge and strengthen throughout adolescence, with older adolescents demonstrating more accurate intuitive responses than younger ones. At the same time, cognitive ability appears to play a more substantial role in both intuitive and deliberative reasoning as adolescents grow older, though the full “smart intuitor” pattern seen in adults does not yet materialize. In other words, while logical principles may tend to become increasingly automated with age, the developmental process is not complete by late adolescence, and it is likely that the influence of cognitive ability on reasoning continues to evolve.

## Acknowledgements

This research was supported by a research grant (INTUIT, ANR-23-CE28-0004-01) from the Agence Nationale de la Recherche, France.

We would like to thank Laurence Berthier and Jean-Luc Berthier for their invaluable help with school recruitment. We would also like to thank John Abi Hana, Jérémie Beucler, Léopold Cayzeele, Cécile Charbit, Philippe Charbit, Anaïs Crepy and Célia Gumbs for their help with data collection.

During the preparation of this work the authors used GPT-4 in order to assist language editing. The authors reviewed and revised the content as needed and take full responsibility for the content of the published article.

## Declaration of Interest Statement

The authors report there are no competing interests to declare.

## References

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 and 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 and Reasoning*, *26*(1), 1–30. https://doi.org/10.1080/13546783.2018.1552194

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*(3). https://doi.org/10.1017/S0140525X07001653

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of Abbreviated Nine-Item Forms of the Raven's Standard Progressive Matrices Test. *Assessment*, *19*(3), 354–369. https://doi.org/10.1177/1073191112446655

Boissin, E., Charbit, L., Borst, G., Caparos, S., & De Neys, W. (in prep). *Testing The Logical Intuition's Automatization Assumption: The Effect of Training For Early And Late Adolescents*

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

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

De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, *46*. https://doi.org/10.1017/S0140525X2200142X

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., & Verschueren, N. (2006). Working Memory Capacity and a Notorious Brain Teaser. *Experimental Psychology*, *53*(2), 123–131. https://doi.org/10.1027/1618-3169.53.1.123

Diedenhofen, B., & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

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

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. (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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

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

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic perspectives, 19(*4), 25-42. https://doi.org/10.1257/089533005775196732

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and Content: The Use of Base Rates as a Continuous Variable. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 513–525. https://doi.org/10.1037/0096-1523.14.3.513

Kahneman, D. (2011). *Thinking, Fast and Slow* (Strauss & Giroux, Eds.; Farrar).

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https://doi.org/10.1037/h0034747

Kassambara, A. (2023a). *ggpubr: "ggplot2" Based Publication Ready Plots*. https://doi.org/10.32614/CRAN.package.ggpubr

Kassambara, A. (2023b). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*. https://doi.org/10.32614/CRAN.package.rstatix

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

Langener, A. M., Kramer, A. W., van den Bos, W., & Huizenga, H. M. (2021). A shortened version of Raven's standard progressive matrices for children and adolescents. *British Journal of Developmental Psychology*, *40*(1), 35–45. https://doi.org/10.1111/bjdp.12381

Lawrence, M. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments*. https://doi.org/10.32614/CRAN.package.ez

McDonald, B. (2002). A Teaching Note on Cook's Distance - A Guideline. *Research Letters in the Information and Mathematical Sciences, 3*, 127–128.

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

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

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

Pennycook, G., Newton, C., & Thompson, V. A. (2022). Base-rate neglect. In R. Pohl (Ed.), *Cognitive illusions* (3rd ed., pp. 17-34). Routledge. https://doi.org/10.4324/9781003154730-5

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

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

Raven, J., Raven, J. C., & Court, J. H. (1998a). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices*. Oxford Psychologists Press; The Psychological Corporation.

Raven, J., Raven, J. C., & Court, J. H. (1998b). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4: the Advanced Progressive Matrices*. Oxford Psychologists Press; The Psychological Corporation.

Revelle, W. (2025). *psych : Procedures for Psychological, Psychometric, and Personality Research.* https://doi.org/10.32614/CRAN.package.psych

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. https://doi.org/10.1037/0033-2909.119.1.3

Stanovich, K. E. (2011). *Rationality and the reflective mind*. Oxford University Press.

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. (1999). Discrepancies Between Normative and Descriptive Models of Decision Making and the Understanding/Acceptance Principle. *Cognitive Psychology*, *38*(3), 349–385. https://doi.org/10.1006/cogp.1998.0700

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

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

Toplak, M. E., & Flora, D. B. (2020). Resistance to cognitive biases: Longitudinal trajectories and associations with cognitive abilities and academic achievement across development. *Journal of Behavioral Decision Making*, *34*(3), 344–358. https://doi.org/10.1002/bdm.2214

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory and Cognition*, *39*(7), 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental Psychology*, *50*(4), 1037–1048. https://doi.org/10.1037/a0034910

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & van den Brand, T. (2024). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://doi.org/10.32614/CRAN.package.ggplot2

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. https://doi.org/10.32614/CRAN.package.dplyr

Wickham, H., Vaughan, D., & Girlich, M. (2024). *tidyr: Tidy Messy Data*. https://doi.org/10.32614/CRAN.package.tidyr

# Supplementary Materials

## A. Bat-and-ball task

The bat-and-ball task was preregistered but excluded from the main analyses following reviewers' recommendations, due to floor-level performance and poor internal consistency, particularly among younger adolescents. The task is presented here as exploratory data.

### Method

#### *Material*

**Bat-and-Ball items.** We selected eight content modified versions of the bat-and-ball problem that were originally adapted in French by Boissin et al. (in prep) from the work of (Raoelison & De Neys, 2019). They were modified versions of the bat-and-ball problem (e.g., "*In a company there are 150 men and women in total. There are 100 more men than women. How many women are there?*"), which used quantities instead of prices (Bago & De Neys, 2019; Janssen et al., 2020; Raoelison & De Neys, 2019). Bat-and-ball problems were presented serially, with the sentences introduced one after the other. Participants had to select the correct response among four response choices which were composed of (1) the correct response (i.e., "5 cents" in the original bat-and-ball), (2) the intuitively cued "heuristic" response (i.e., "10 cents"), (3) a foil option which was the sum of correct and heuristic answers (i.e., "15 cents"), and (4) a second foil option which was the second greatest common divisor (i.e., "1 cent").

Mathematically speaking, the correct equation to solve the standard bat-and-ball problem is: "$1.00 + 2x = $1.10", instead, people are thought to be intuitively using the "$1.00 + x = $1.10" equation to determine their response (Kahneman, 2011). The latter equation was used to determine the "heuristic" answer option, and the former to determine the correct answer option for this problem. The four response choices appeared in a random order. For instance:

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

- *25*
- *50*
- *75*
- *10*

Here the correct answer is 25.

Half of the problems were featured in their standard "conflict" version and the other half in their no-conflict version. In these control no-conflict problems, we deleted the critical relational "more than" statement. The heuristic intuition thus cued the correct response (De Neys et al., 2013; Travers et al., 2016). We presented the same four answer options as for a corresponding standard conflict version. We added three words to the control problem questions (e.g., "How many women are there in the office?") in order to equate the semantic length of the conflict and no-conflict (control) versions (Raoelison & De Neys, 2019). For instance:

*In a company, there are 150 men and women in total.*
*There are 100 men. How many women are there in the office?*

- *25*
- *50*
- *75*
- *10*

Two sets of problems were used to counterbalance problem content. The conflict problems in one set were the no-conflict problems in the other, and vice-versa. As a result, set A contained four conflict and four no-conflict items, while set B contained the matching four no-conflict and four conflict items. Participants were randomly assigned to one of the two sets. The presentation order of the problems was randomized for each participant. In total, each participant solved four conflict and four no-conflict items.

**Two-response format.** As in the base-rate task, participants responded to each bat-and-ball problem using a two-response paradigm, where they first provided a ‘fast’ answer, directly followed by a second ‘slow’ answer (Thompson et al., 2011). The initial response deadline was set at 5 seconds, based on pretesting by Bago and De Neys (2019). The allotted time corresponded to the time required to read the problem, the question and answer alternatives, move the mouse, and select an answer among the possibilities (see Bago & De Neys, 2019, for details). Furthermore, participants are also under a secondary task load when giving their initial response, making intuitive responding even more challenging. Obviously, the time limit and cognitive load were applied only for the initial response, and not for the final one where participants were allowed to deliberate (see below).

**Two-response format and development.** The two-response paradigm can be quite challenging, especially for younger adolescents, who may have difficulty reading and comprehending the problems under strict time constraints and cognitive load. However, several studies (Raoelison et al., 2021; Boissin et al., in prep) have shown that we can use the

two-response procedure with young and old adolescents on various reasoning tasks. In these studies, the same deadline was used for both younger (around 12 years old) and older adolescents (around 17 years old) when contrasting the intuitive reasoning performance of these age groups. Although both the deadline and the load memorization task were challenging, (younger) participants were able to meet them on the vast majority of trials. Similarly, in the current study, (younger) participants met constraints on most trials (see Trial exclusions below). Moreover, no-conflict accuracies for initial responses were very high among younger adolescents (*M* = 88.96%, *SD* = 11.52%), indicating that 7th graders were able to solve bat-and-ball problems under these stringent conditions and refrained from mere random guessing. These results suggest that the two-response procedure can be effectively used with the bat-and-ball task in a young adolescent sample.

***Procedure***

As shown in Figure S1, bat-and-ball trials followed a structure similar to base-rate trials. Each trial began with a fixation cross displayed for 2000 ms, followed by the first sentence of the problem (e.g., "*In a company, there are 150 men and women in total.*") for another 2000 ms. Next, a target pattern for the memorization task appeared for 2000 ms. Participants then saw the full problem, which included the second sentence and the question (e.g., "*There are 100 more men than women. How many women are there?"),* along with four answer choices. They had 5000 ms to select an initial response. After 3000 ms, the screen background turned yellow to indicate that time was running out.

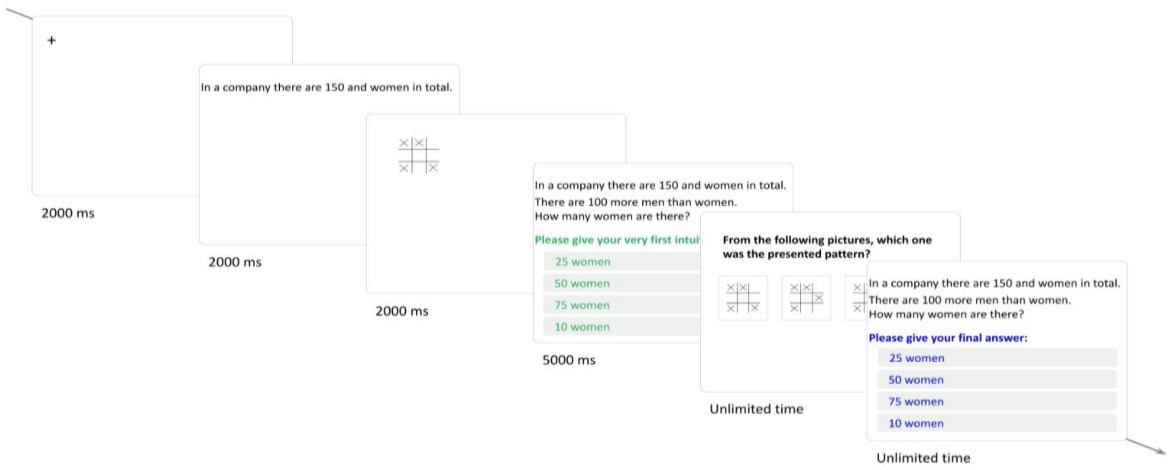


**Figure S1.** Time course of a complete two-response trial for the bat-and-ball task.

## Results

### *Trial exclusion*

We discarded trials in which participants either failed to respond before the deadline or did not successfully complete the load memorization task, as we could not guarantee that their initial responses on these trials were free from deliberation (Bago & De Neys, 2017). Table S1 details the excluded trials for the bat-and-ball task by age group. Ultimately, we analyzed 79.81% of bat-and-ball trials for older adolescents and 70.28% for younger adolescents.

**Table S1.** Percentage of excluded and analyzed trials by task and age group.

| Age group | Excluded trials | | Analyzed trials (%) |
|---|---|---|---|
| | Missed deadline (%) | Failed load for the remaining trials (%) | |
| Younger adolescents | 6.63 | 23.08 | 70.28 |
| Older adolescents | 4.74 | 15.45 | 79.81 |

**Note.** Excluded trials were calculated in two steps. First, the trials with a missed deadline were discarded. Next, from the remaining trials, those with a failed load were excluded.

We also measured the average individual contribution. For the bat-and-ball task, on average older adolescent contributed 2.94 (*SD* = 1.04) conflict trials out of four and 3.16 (*SD* = 0.87) no-conflict trials out of four; younger contributed 2.63 (*SD* = 1.11) conflict trials, and 2.63 (*SD* = 1.06) no-conflict trials.

### *Bat-and-Ball familiarity*

Following Haigh (2016), we screened participants for prior familiarity with the original bat-and-ball problem as previous studies have shown that individuals who are familiar with the Cognitive Reflection Test (CRT, Frederick, 2005)—a brief three-item questionnaire that includes the bat-and-ball problem—tend to perform better on the problem than their naive counterparts (Bialek & Pennycook, 2018; Haigh, 2016; Stieger & Reips, 2016). In our study, 54 younger adolescents reported having encountered the bat-and-ball problem before, but only two provided the correct answer ("5 cents"). Among the older adolescents, 97 reported prior exposure to the problem, and 31 of them gave the correct answer. Critically, individuals with higher cognitive ability may encounter the bat-and-ball problem more frequently, giving them more opportunities to automatize and generate correct intuitive responses (Bialek &

Pennycook, 2018). If we do not account for this, it could artificially inflate support for the 'smart intuitor' hypothesis.

Thus, as preregistered, we conducted two sets of analyses: one with and one without excluding the two younger and 31 older adolescents who reported both prior exposure and responded correctly.

## Reasoning performance

### *Accuracy*

Bat-and-ball conflict accuracies ranged from 0% to 100% in both age groups. Nearly all participants scored 0% correct on conflict items, both for initial and final responses (see Figure S6, for distribution plots). Figure S2 shows that older adolescents outperformed younger adolescents at both the initial and final response stages. Older adolescents descriptively tended to improve on conflict items from the initial (*M* = 6.96%, *SD* = 7.22%) to the final response stage (*M* = 9.34%, *SD* = 4.17%) while younger adolescents hardly showed an improvement (initial *M* = 3.90%, *SD* = 3.97%; final: *M* = 4.06%, *SD* = 3.11%). The ANOVA between age group and response stage revealed a significant main effect of age group, $F(1, 310) = 4.18$, $p = .042$, $\eta^2_G = .011$, but no significant main effect of response stage $F(1, 310) = 1.45$, $p = .229$, $\eta^2_G = .001$, nor a significant interaction between age group and response stage $F(1, 310) = 1.10$, $p = .294$, $\eta^2_G = .001$. These results suggest that (1) deliberation did not significantly improve performance, and (2) unlike in the base-rate task, older adolescents did not gain more from deliberation than their younger counterparts. Note that, Figure S2 also shows that accuracy was lower for the bat-and-ball problems compared to the base-rate problems, with performance nearing floor level.

For the no-conflict problem, Figure S2 further shows that accuracy on the bat-and-ball no-conflict problems was high for both older (initial: *M* = 91.93%, *SD* = 7.59%, final: *M* = 98.02%, *SD* = 1.96%) and younger participants (initial: *M* = 88.96%, *SD* = 11.52%; final: *M* = 97.40%, *SD* = 3.55%). T-tests confirmed that both age groups' accuracy exceeded chance (i.e., 25%) for both initial and final responses, all $p < .001$. Similar to the base-rate task, this confirms that participants read and processed the material and did not engage in random guessing on the bat-and-ball task.

### *Direction of change*

For the direction of change of bat-and-ball problems, Figure 3 shows that both older and younger adolescents primarily provided 00 response patterns (00: *M* = 92.86%, *SD* = 20.97% for younger adolescents, and *M* = 88.40%, *SD* = 28.56% for older adolescents). Similarly to the base-rate direction of change analysis, Figure S2 shows that older adolescents tend to

give fewer 00 response patterns than younger adolescents, $t(310) = 1.57$, $p = .118$, $d = .18$, and significantly more 11 responses $t(310) = -2.51$, $p = .013$, $d = -.28$. The analysis revealed no significant difference in the 01 proportion between younger and older adolescents, $t(310) = -0.77$, $p = .386$, $d = -0.10$. This suggests that, for the bat-and-ball task, while participants do not show significant improvement after deliberation with age, they do improve in producing correct intuitions and require less deliberation than the younger participants.

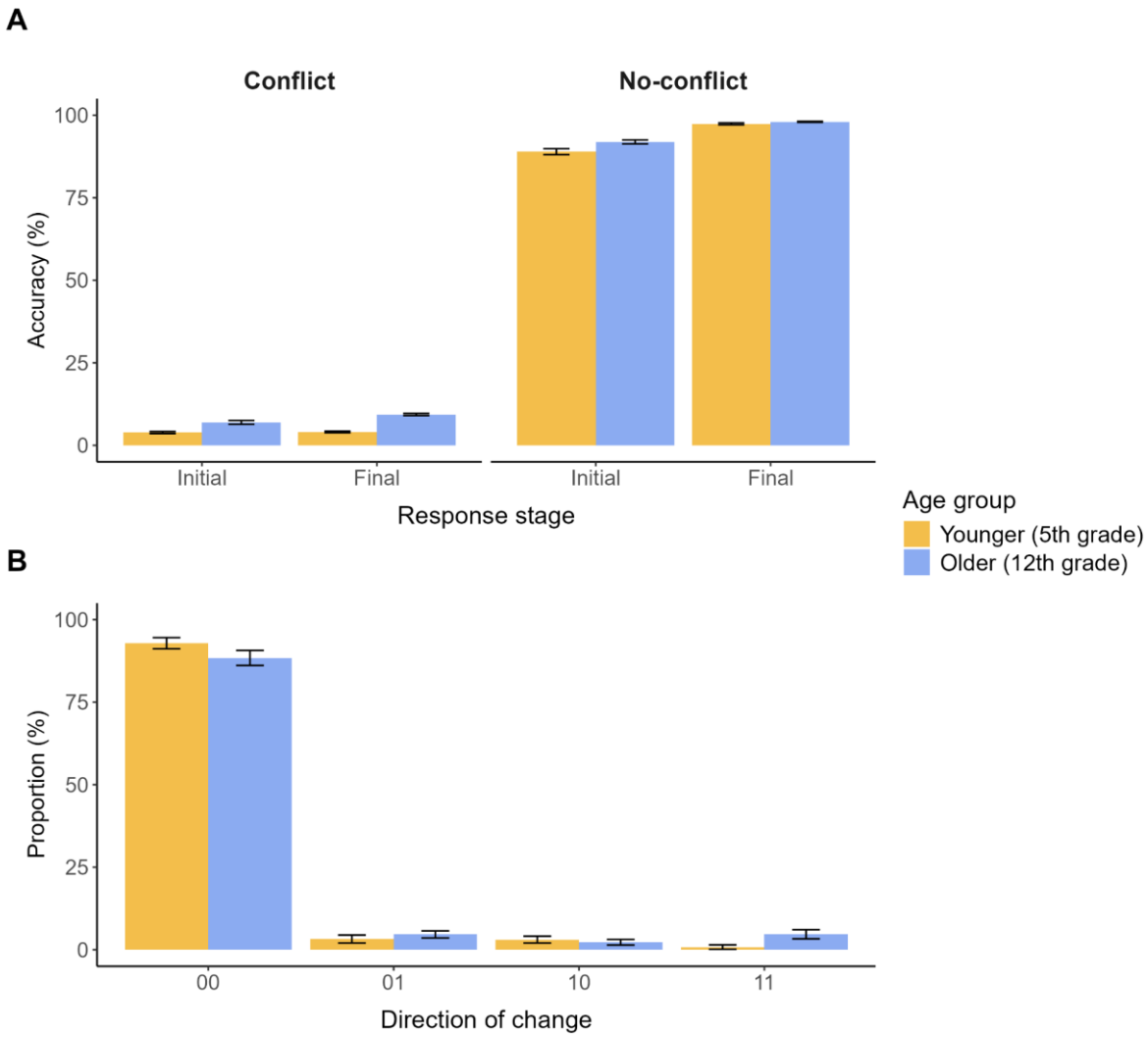


**Figure S2.** Accuracy and Proportions of direction of change for the bat-and-ball task. Error bars represent Standard Error of the Mean (SEM). (A) Mean response accuracy for conflict items and no-conflict items as a function of response stage. (B) Proportions of direction of change (i.e., 00 = both 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) categories for conflict items.

***Reasoning performance with bat-and-ball familiarity exclusions***

For these analyses, we excluded the two younger and 31 older adolescents who both reported prior exposure and responded correctly to the classic bat-and-ball task. The conclusions remain similar to those obtained without these exclusions and are reported below.

**Accuracy.** When participants familiar with the task were excluded, mean accuracy for both younger and older adolescents were below 5%, regardless of response stage. As figure S3 shows, older adolescents tended to deteriorate slightly from the initial ($M$ = 4.04%, $SD$ = 4.65%) to the final response stage ($M$ = 3.65%, $SD$ = 2.90%), whereas younger adolescents tended to improve slightly from the initial ($M$ = 3.95%, $SD$ = 4.03%) to the final response stage ($M$ = 4.11%, $SD$ = 3.15%). The ANOVA between age group and response stage revealed no significant main effect of age group, $F(1, 278) = 0.01$, $p = .906$, $\eta^2_G = .000$, and no significant main effect of response stage, $F(1, 278) = 0.01$, $p = .915$, $\eta^2_G = .000$, nor a significant interaction between age group and response stage, $F(1, 278) = 0.07$, $p = .794$, $\eta^2_G = .000$. These results suggest that both older and younger adolescents performed poorly and did not benefit from deliberation.

For the no-conflict problems, Figure S3 further shows that accuracy was high for both older (initial: $M$ = 92.44%, $SD$ = 6.88%, final: $M$ = 98.32%, $SD$ = 2.43%) and younger participants (initial: $M$ = 88.82%, $SD$ = 11.69%; final: $M$ = 97.37%, $SD$ = 3.61%). T-tests confirmed that accuracy of both age groups exceeded chance (i.e., 25%) in initial and final response, all $p < .001$. This confirms that participants read and processed the material and did not engage in random guessing in the bat-and-ball task.

**Direction of change.** For the direction of change on bat-and-ball problems, Figure S3 shows that both older and younger adolescents primarily provided 00 response patterns, i.e., incorrect responses at both response stages (00: $M$ = 92.76%, $SD$ = 21.09% for younger adolescents, and $M$ = 93.95%, $SD$ = 19.67% for older adolescents). Both age groups provided very few 01 response patterns, i.e., correct response only at the deliberate stage (01: $M$ = 3.29%, $SD$ = 14.99% for younger adolescents, and $M$ = 2.02%, $SD$ = 8.67% for older adolescents). The analysis revealed no significant difference in the 01 proportion between younger and older adolescents, $t(278) = 0.85$, $p = .397$, $d = 0.10$. Both age groups also provided very few 11 patterns, i.e., correct responses at both stages (11: $M$ = 0.82%, $SD$ = 8.35% for younger adolescents, and $M$ = 1.63%, $SD$ = 10.46% for older adolescents). The analysis revealed no significant difference in the 11 proportion between younger and older adolescents, $t(278) = -0.72$, $p = .474$, $d = -0.09$. These findings suggest that participants do not improve with age in generating either corrective deliberations or correct intuitions.

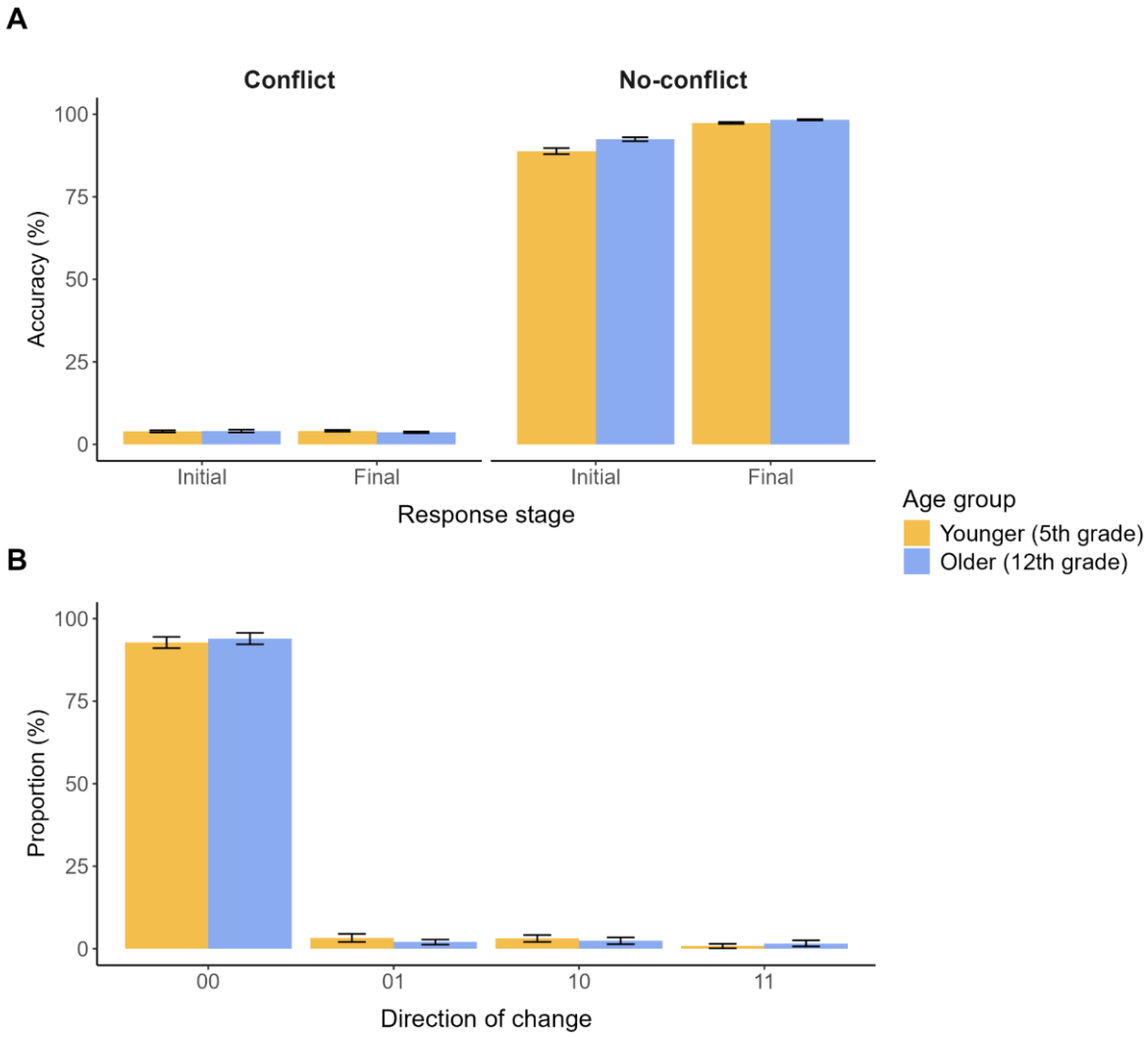


**Figure S3.** Accuracy and Proportions of direction of change for the bat-and-ball task with familiarity exclusions. Error bars represent Standard Error of the Mean (SEM). (A) Mean response accuracy for conflict items and no-conflict items as a function of response stage. (B) Proportions of direction of change (i.e., 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) categories for conflict items.

### Cognitive ability correlations

We evaluated the impact of potential outliers on our correlation analyses by computing Cook's distance for each data point, using a cutoff of 0.7 as recommended by McDonald (2002). No observations exceeded this threshold (all $D < 0.7$), indicating that our results were not driven by any single data point.

***Accuracy correlations***

Table S2 shows the correlations between cognitive ability and reasoning accuracy on the bat-and-ball task, both without and with familiarity exclusions.

When familiarity exclusions were not applied, older adolescents showed positive correlations between cognitive ability and both final, $r(156) = 0.30$, $p < .001$, and initial reasoning accuracy, $r(156) = 0.24$, $p = .002$. However, when participants familiar with the bat-and-ball task were excluded, older adolescents' cognitive ability no longer correlated with either final accuracy, $r(126) = 0.12$, $p = .181$, or initial accuracy, $r(126) = 0.05$, $p = .592$.

In contrast, younger adolescents showed no positive correlation between cognitive ability and reasoning accuracy at either the final, $r(152) = -0.02$, p = .774, or the initial stage, $r(152) = -0.19$, $p = .020$. This lack of positive correlation with cognitive ability was consistent when excluding participants familiar with the task, both at the final, $r(150) = -0.02$, $p = .809$, and the initial stage, $r(150) = -0.19$, $p = .022$. If anything, younger adolescents showed a negative correlation for initial responses, suggesting that those with lower cognitive ability were more likely to select correct responses, possibly reflecting increased guessing among participants lower in cognitive capacity.

**Table S2.** Correlations between cognitive ability and bat-and-ball accuracy at the initial and final response stages for conflict items.

| Task | Age group | Response stage | *r* | *p* | df |
|---|---|---|---|---|---|
| Bat-and-Ball | Younger | Initial | -0.19* | .020 | 152 |
| | | Final | -0.02 | 0.774 | 152 |
| | Older | Initial | 0.24* | .002 | 156 |
| | | Final | 0.30*** | <.001 | 156 |
| Bat-and-Ball with | Younger | Initial | -0.19* | .022 | 150 |
| familiarity exclusions | | Final | -0.02 | .809 | 150 |
| | Older | Initial | 0.05 | 0.592 | 126 |
| | | Final | 0.12 | .181 | 126 |

* $p < .05$; ** $p < .01$; *** $p < .001$.

***Direction of change correlations***

To disentangle the relationship between cognitive ability and sound intuition from its relationship with corrective deliberation, we examined correlations between cognitive ability and the 01 (i.e., correct response only at the deliberate stage) and 11 (i.e. correct response from the intuitive stage) patterns. Table S3 presents these key correlations for bat-and-ball task, both without and with familiarity exclusion.

Older adolescents' cognitive ability correlated positively with the proportion of both 01 responses, $r(156) = 0.17$, $p = .031$, and 11 responses, $r(156) = 0.32$, $p < .001$. The difference between the two correlations was marginally significant , $z = -1.95$, $p = .051$. This suggests that for older adolescents, cognitive ability tends to be more strongly linked to correct intuitive responding than correct deliberative responding. However, when participants familiar with the bat-and-ball task were excluded, older adolescents' cognitive ability no longer correlated with either 01, $r(126) = 0.16$, $p = .078$, or 11 responses, $r(126) = 0.03$, $p = .733$.

In contrast, younger adolescents showed no positive correlation between cognitive ability and 01 responses, $r(152) = -0.04$, $p = .606$, nor between cognitive ability and 11 responses, $r(152) = 0.03$, $p = .734$. This lack of positive correlation with cognitive ability was consistent when excluding participants familiar with the task, both for 01, $r(150) = -0.04$, $p = .635$, and 11 responses, $r(150) = 0.03$, $p = .719$.

**Table S3.** Correlations between cognitive ability and bat-and-ball direction of change categories for conflict items.

| Task | Age group | Direction of change category | *r* | *p* | df |
|---|---|---|---|---|---|
| Bat-and-Ball | Younger | 01 | -0.04 | .606 | 152 |
| | | 11 | 0.03 | .734 | 152 |
| | Older | 01 | 0.17* | .031 | 156 |
| | | 11 | 0.32*** | <.001 | 156 |
| Bat-and-Ball with | Younger | 01 | -0.04 | .635 | 150 |
| familiarity exclusions | | 11 | 0.03 | .719 | 150 |
| | Older | 01 | 0.03 | .733 | 126 |
| | | 11 | 0.16 | .078 | 126 |

* $p < .05$; ** $p < .01$; *** $p < .001$.

# B. Distribution plots

### Raven Matrices

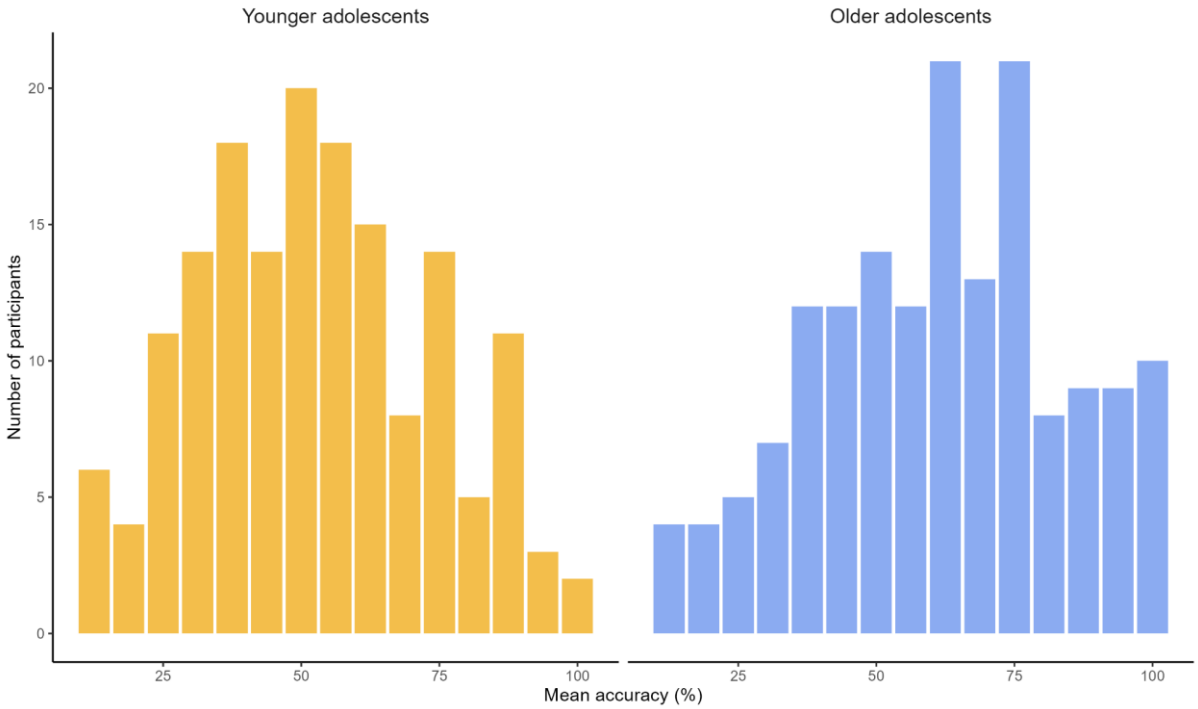


**Figure S4.** Histogram of participants' accuracy on the Raven task for both age groups.

**Base-Rate task**

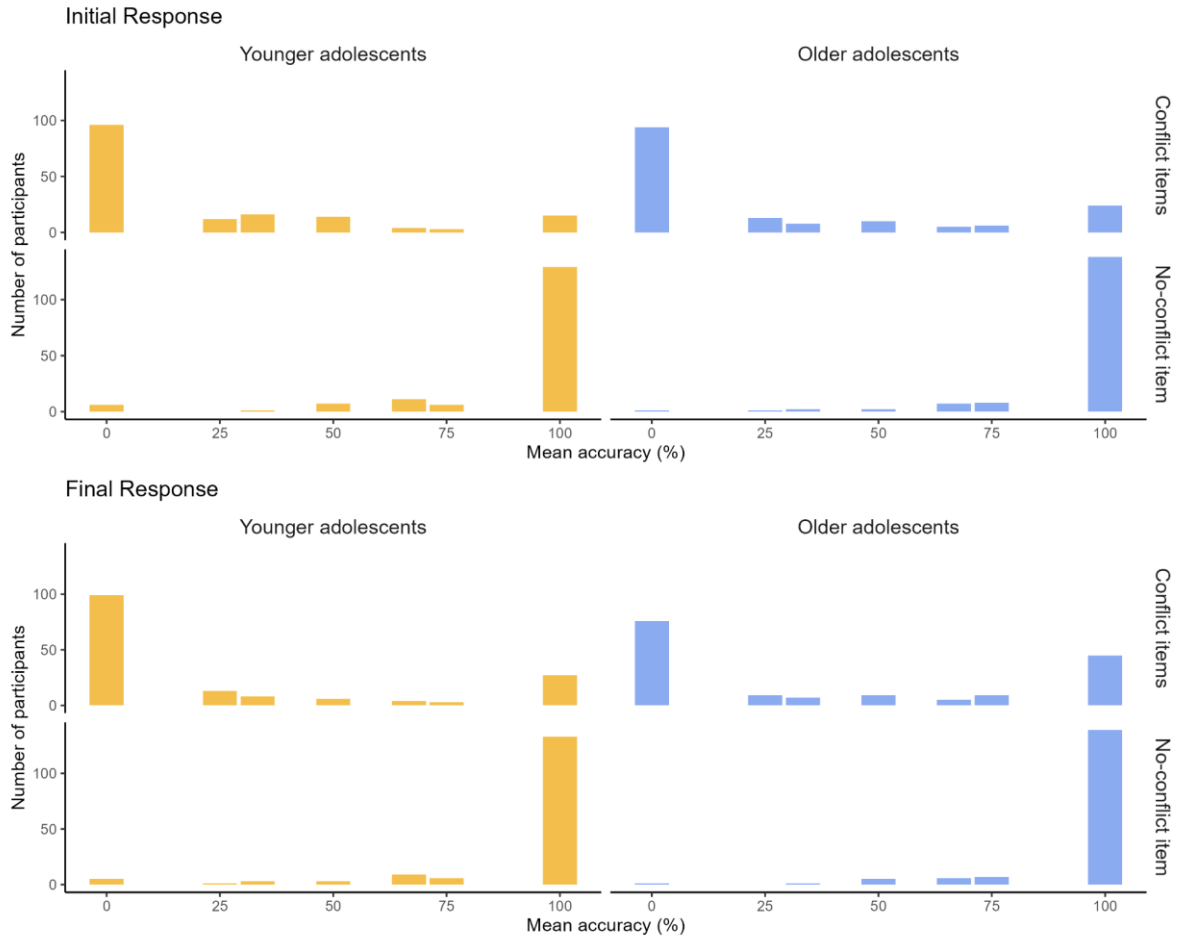


**Figure S5.** Histograms of participants' accuracy on the base-rate task for initial and final responses for both age groups.

**Bat-and-ball task**

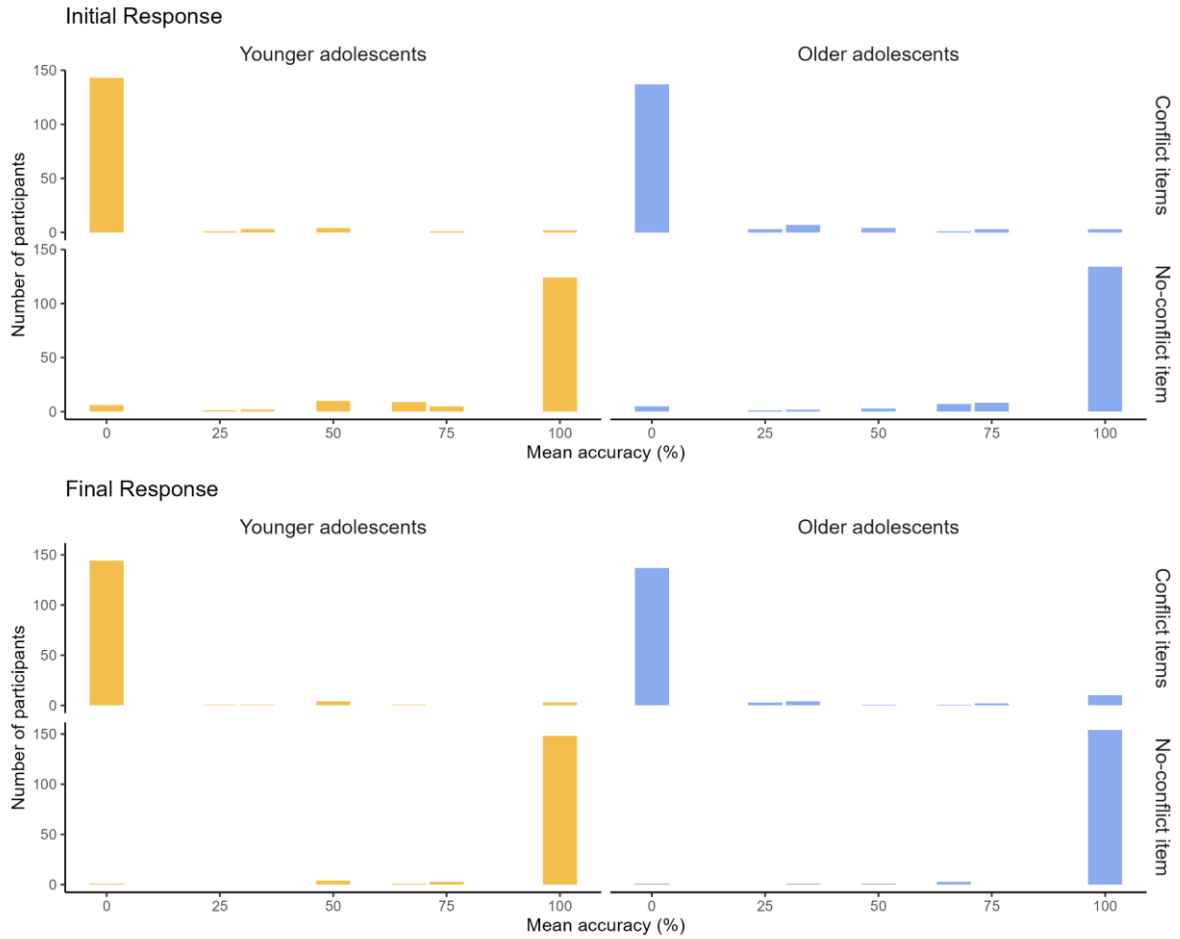


**Figure S6.** Histograms of participants' accuracy on the bat-and-ball task for initial and final responses for both age groups.

## References

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

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

Haigh, M. (2016). Has the standard cognitive reflection test become a victim of its own success? *Advances in Cognitive Psychology*, *12*(3), 145–149. https://doi.org/10.5709/acp-0193-5

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*. https://doi.org/10.1016/j.actpsy.2020.103042

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. https://doi.org/10.1017/S1930297500003405

Stieger, S., & Reips, U. D. (2016). A limitation of the Cognitive Reflection Test: Familiarity. *PeerJ*, *2016*(9). https://doi.org/10.7717/peerj.2395

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