Reactivating Logic? Age-Dependent Effects of Intuitive Debiasing Training

Esther Boissin[1*], Laura Charbit[2], Matthieu Raoelison[2], Grégoire Borst[2], Serge Caparos[3,4] & Wim De Neys[2,5]

[1] Department of Psychology, Cornell University, USA

[2] LaPsyDE, University of Paris Cité, France

[3] DysCo Lab, University of Paris 8, France

[4] Institut Universitaire de France, France

[5] CNRS, France

*Corresponding author: [boissinesther@gmail.com], Cornell University, Uris Hall, 211, 109 Tower Rd, Ithaca, NY 14853

In press at Thinking & Reasoning

June 30th, 2026

Abstract

Previous work has shown that short debiasing interventions in which the problem solution is explained can boost adults' intuitive reasoning. This effect has been attributed to the reactivation of prior logico-mathematical knowledge acquired through schooling. However, this hypothesis has not yet been directly tested. To address this issue, we compared the impact of debias training on two classic reasoning tasks (i.e., the bat-and-ball and base-rate problems) on reasoning performance at the end (11-12th grade) and start of (7^{th}-8^{th} grade) secondary education (i.e., a group with a more and less developed logico-mathematical knowledge base). We adopted a two-response paradigm to distinguish between intuitive and deliberate responses before and after the training. Across both tasks, training improved performance (compared to a no-training control group) in both age groups, primarily through increased intuitive accuracy. Critically, debiasing was more likely to produce logical intuitions in late than early adolescents. These findings suggest that the success of intuitive boosting relies on the extent to which relevant logical principles have been practiced to the point of automaticity.

Keywords: debiasing training, intuition, adolescent, dual-process

Introduction

Influential research in reasoning and decision-making has shown that individuals routinely violate basic logical and mathematical principles. Rather than engaging in effortful reflection, people often rely on fast, intuitive impressions to solve problems. These intuitions can be useful—because they operate quickly and with minimal cognitive effort—but they also frequently generate responses that conflict with logical or probabilistic principles. This bias can be illustrated by Frederick's (2005) well-known bat-and-ball problem: "*A bat and a ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?*"

Intuitively, many reasoners promptly conclude that the ball costs "10 cents". This response is given in a majority of cases (Frederick, 2005; Toplak et al., 2014), even in samples composed of highly qualified university students (Bourgeois-Gironde & Van der Henst, 2009; Frederick, 2005), and even after repeated exposure to the problem (Raoelison & De Neys, 2019; Stagnaro et al., 2018). However, although it is intuitively appealing, the answer is not correct. If the ball costs 10 cents, and the bat costs $1 more, then the bat would cost $1.10. If the bat costs $1.10,

then the total would be $1.20 and not $1.10 as stated. On reflection, the ball costs 5 cents and the bat—which costs $1 more—costs $1.05.

Traditionally, such biased responses are explained by dual-process theories, which distinguish between an intuitive system (System 1) and a deliberative system (System 2; Kahneman, 2011). According to this model, errors arise from overreliance on the fast, intuitive system and insufficient engagement of the slower, resource-intensive deliberative system. In contrast, correct responses require the deliberate correction of an initial intuitive error (Evans & Stanovich, 2013; Kahneman & Frederick, 2002; Morewedge & Kahneman, 2010). However, because most reasoners tend to minimize cognitive effort (Kahneman, 2011), they default to the intuitive system and often accept the first response that comes to mind without considering that it might be incorrect.

Nevertheless, recent studies have challenged the traditional view by showing that accurate responses can also arise intuitively without deliberate correction (e.g., Bago & De Neys, 2017, 2019; Newman et al., 2017; Thompson et al., 2011). These studies adopted a two-response paradigm (Thompson et al., 2011) in which participants are asked to give two consecutive responses to a reasoning problem. First, they respond as fast as possible with the first intuitive hunch that comes to mind. Next, they can take all the time they need to reflect on the problem and give a final response. To make maximally sure that the initial response is generated intuitively, the initial intuitive response needs to be given under time pressure and/or cognitive load (Bago & De Neys, 2017; Newman et al., 2017). Results have shown that sound reasoners often give a correct response as soon as the initial intuitive stage (Bago & De Neys, 2017, 2019; Newman et al., 2017; Raoelison & De Neys, 2019; Thompson et al., 2011), supporting the idea that correct responses do not always require costly deliberation. However, while intuitive correct responses do occur, they remain relatively rare, and the majority of reasoners still fall prey to biases (Bago & De Neys, 2017, 2019; Burič & Šrol, 2020; Janssen et al., 2020; Newman et al., 2017).

Given this tendency toward biased responses, considerable research has focused on developing interventions aimed at debiasing reasoning and promoting correct responses (e.g., Lilienfeld et al., 2009; Milkman et al., 2009; Nisbett, 1993). Some interventions have proven successful in that respect. For instance, brief explanations of the typical bias and correct solution strategy often enable reasoners to respond accurately on similar tasks (Boissin et al., 2021, 2022, 2023; Claidière et al., 2017; Franiatte et al., 2024a, 2024b; Hoover & Healy, 2021; Morewedge et al., 2015; Purcell et al., 2021; Trouche et al., 2014). Notably, such training frequently leads to correct responses generated in the initial, intuitive response stage (Boissin et al., 2021, 2022, 2023). These findings suggest that intuitive processes, or “System 1,” can be trained to avoid

bias. Once individuals understand a problem's solution, the correct response can become sufficiently dominant that no corrective deliberative step is needed; instead, reasoners may generate the correct answer directly from the outset. This supports the idea that training can shift reasoners from biased to sound intuiting (i.e. "System 1" debiasing Boissin et al., 2021, 2022, 2023; Franiatte et al., 2024a, 2024b).

A key question concerns the mechanisms underlying the efficiency of the short training intervention. The "logical intuition" account argues that the training may reactivate already-acquired knowledge which is necessary to solve the problem (Boissin et al., 2021, 2022). That is, people have—to some extent at least—already automatized the computation of basic mathematical and logical principles through repeated educational exposure and practice (see De Neys & Pennycook, 2019). People may nevertheless be biased because the activation of such correct intuitions might be outcompeted by conflicting incorrect intuitions. In other words, implicit knowledge about the correct logical solution strategy may be there already, and people simply need to be reminded how to apply it (Boissin et al., 2021, 2022). By highlighting the relevance of the logical principles, the training may boost the activation strength of the logical intuitions so that they outcompete conflicting incorrect intuitions. Consequently, adults may benefit from the training precisely because they already possess logical knowledge that has been instantiated to some extent. However, the link between automatization of prior logical knowledge and intuitive trainability has not been investigated yet.

Some indirect evidence supports the above hypothesis by showing that younger adolescents (i.e., 7th graders)—who had less opportunity to automatize the relevant logico-mathematical principles during their school years (De Neys, 2023)—generate fewer intuitive correct responses than older adolescents (i.e., 12th graders; Raoelison et al., 2021). The performance difference between age groups may thus reflect varying degrees of exposure to these principles. If intervention effectiveness indeed depends on reasoners' pre-existing logical knowledge, intervention success should increase with age, and the difference between age groups should be larger at the intuitive level.

In the current study, we investigated the impact of prior logico-mathematical knowledge on the training of correct intuitions by comparing the training effect across adolescents at the start (early: 7th–8th grade) and end (late: 11th–12th grade) of secondary education, using a two-response paradigm. We used both bat-and-ball problems (Study 1) and base-rate problems (Study 2; Kahneman & Tversky, 1973) to assess reasoning performance and test the generalizability of intuitive debiasing. Both tasks involve distinct logical principles (i.e., solving equations with unknowns; integrating base rates into probability judgments) that are taught and

practiced throughout the secondary school curriculum. In each study, we compared pre- and post-training performance within each age group and contrasted it with that of a no-training control group. We hypothesized that late adolescents, having received more extensive exposure to logical principles, would benefit more from the training than early adolescents. Critically, this age-related difference was expected to be most pronounced at the intuitive (i.e., initial) response stage.

Material and Methods

Studies 1 and 2 were conducted separately and preregistered independently to examine the developmental effects of debiasing training across different reasoning tasks. Given their similar designs and effects, we present them in parallel within a unified methods and results section.

Preregistration and data availability

The study design and research questions were preregistered separately for each task on the AsPredicted website (https://aspredicted.org) and is stored on the Open Science Framework. No specific analyses were preregistered. All data and materials are also available on the Open Science Framework (https://osf.io/96avw/?view_only=f0ae615c9ac04c1c899d3fab640129dd).

Participants

In Study 1 using bat-and-ball items, we recruited 242 French pupils: 116 early adolescents (7th–8th grade; 61 girls, 6 non-binary; M age = 12.7 ± 0.1 years) and 126 late adolescents (11th–12th grade; 71 girls, 5 non-binary; M age = 16.5 ± 0.1 years). In Study 2 using base-rate items, we recruited 220 French pupils: 116 early adolescents (59 girls, 4 non-binary; M age = 12.8, SEM = 0.6) and 104 late adolescents (53 girls, 5 non-binary; M age = 16.3, SEM = 0.6). In both studies, participants were randomly assigned to a control or training group. In Study 1, 60 early adolescents were assigned to the control group and 56 to the training group; among late adolescents, 56 were in the control group and 70 in the training group. In Study 2, 59 early adolescents were assigned to the control group and 57 to the training group; among late adolescents, 51 were in the control group and 53 in the training group. Sample sizes were based on Boissin et al.'s (2021, 2022) training studies, which included approximately 50 participants per group.

Following Boissin et al. (2021), participants in Study 1 were screened during the intervention block for prior familiarity with the bat-and-ball problem. Twenty late adolescents reported prior knowledge and gave the correct answer (“5 cents”). In previous work (e.g., Bago & De Neys, 2019), such participants have been excluded to control for prior knowledge effects. However, similar findings have been reported regardless of whether these participants are included (Franiatte et al., 2024b; Raoelison & De Neys, 2019), and prior exposure has been shown not to affect subsequent responses (Meyer et al., 2018). This was the case in our study: including or excluding them did not affect the pattern of results. As preregistered, we retained these participants in the main analyses and reported the results excluding them in Supplementary Material Section A.

Materials

Each study followed the same structure with three blocks: a pre-intervention block, an intervention, and a post-intervention block. The pre- and post-intervention blocks each included four conflict, two transfer-neutral, and four no-conflict problems. During the intervention, participants completed two additional conflict problems. Participants in the training group received a brief explanation of the correct solution after each item; those in the control group received no explanation. In total, each participant completed 22 problems.

Pre- and post-intervention block

Conflict and No-conflict items

Study 1: Bat-and-ball items. In Study 1, during the pre- and the post-intervention block, we presented problems taken from Raoelison and De Neys (2019) and translated from English to French. These were modified versions of the classic bat-and-ball problem, using quantities instead of prices (Bago & De Neys, 2019; Janssen et al., 2020; Raoelison & De Neys, 2019). An example problem might read: ‘In a company there are 150 men and women in total. There are 100 more men than women. How many women are there?’ Participants were asked to select the correct response from four response choices: (1) the correct response (i.e., “5 cents” in the original bat-and-ball), (2) the intuitively cued “heuristic” response (i.e., “10 cents” in the original bat-and-ball), (3) a foil option corresponding to the sum of correct and heuristic answers (i.e., “15 cents”), and (4) a second foil option, calculated as 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?
o 25
o 50
o 75
o 10

Half of the problems were presented in their standard 'conflict' version, in which the intuitively cued 'heuristic' response conflicts with the correct one. To ensure that participants were engaged in the task, we also included 'no-conflict' control problems. In these control problems, we removed the critical relational 'more than' statement, so that the heuristic intuition aligned with the correct response (De Neys et al., 2013; Travers et al., 2016). For example:

In a company, there are 150 men and women in total.
There are 100 men. How many women are there in the office?
o 25
o 50
o 75
o 10

In this example, the intuitively cued '50' answer is also the correct one. 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) (Raoelison & De Neys, 2019). Overall, these control problems should be easy to solve. If participants are paying minimal attention to the task and refrain from random guessing, they should show high accuracy (Bago & De Neys, 2019).

Study 2: Base-rate items. In Study 2, during the pre- and the post-intervention blocks, participants were presented with base-rate problems translated from English to French. Half of these problems were taken from Raoelison et al. (2021) who pre-tested the material to ensure its suitability for earlier adolescents. These items were originally developed by Pennycook et al. (2014). The

remaining problems, also sourced from Pennycook et al. (2014), were newly translated for this study and pre-tested at the end of our main reasoning study (see below).

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 task instructions stressed that the person was drawn randomly from the specified sample.

The problem presentation format was based on Pennycook et al.'s (2014) rapid-response paradigm. 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 businessmen and firemen"). Next, under the first sentence (which remained on the screen) we presented the descriptive information (e.g., Person 'K' is brave). The descriptive information specified a neutral name ('Person K') and a single word personality trait (e.g., "brave") that was designed to trigger the stereotypical association. Finally, participants received the base rate probabilities. As in Pennycook et al., base rates varied between 995/5, 996/4, and 997/3. The following illustrates the full problem format:

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

Pennycook et al. (2014) pre-tested the material to make sure that words that were selected to cue a stereotypical association consistently did so but avoided extremely diagnostic cues. 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., Barbey & Sloman, 2007; Gigerenzer et al., 1988) 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). Half of the problems were presented in their standard 'conflict' version, in which the intuitively cued 'heuristic' response conflicts with the correct one. To ensure that participants were engaged in the task, we also included 'no-conflict' control problems. In the conflict items, the base rate probabilities and the stereotypical information cued conflicting responses (see example above). In the no-conflict items, they both cued the same response (i.e., the description triggered a stereotypical trait of a member of the largest group). The following is an example of a no-conflict problem:

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

These control problems should be easy to solve. If participants are paying minimal attention to the task and refrain from random guessing, they should show high accuracy (Bago et al., 2019). Two sets of 16 unique items (8 pre-intervention and 8 post-intervention block items) were used for counterbalancing purposes.

Following De Neys and Vanderputte (2011), we asked participants at the end of the study to rate the items from Pennycook et al. (2014) that had not previously been pre-tested by Raoelison et al. (2021). This allowed us to check that it cued the intended stereotypes and beliefs across both age groups. Participants indicated their agreement with statements on a scale from -5 to +5. Each pair of statements associated an adjective with two groups (e.g., "A clown is funny" and "An accountant is funny") and was presented simultaneously. Participants rated eight pairs of statements, corresponding to the eight conflict and no-conflict items in the reasoning task (for a total of 16 items). Results confirmed that the materials functioned as intended: in both age groups, the statements with intended stereotypical associations were rated consistently higher than their contrasting counterparts, with minimal variability across age groups (see Supplementary Material, Section B, for details).

Counterbalancing. In both studies, to counterbalance problem content, we created two sets of problems: conflict problems in one set were the no-conflict problems in the other, and vice versa. Consequently, none of the pre- and post-intervention problem contents was repeated within-

subjects. The presentation order of conflict and no-conflict problems was randomized within each set, and participants were randomly assigned to one of the two sets for each block.

Transfer-neutral items

Study 1: Bat-and-ball items. In Study 1, in addition to the bat-and-ball problems, we used another type of reasoning problems to test whether the 'bat-and-ball training' effect could transfer to untrained problems. Previous training work failed to observe such a transfer effect (Boissin et al., 2021, 2022). Our primary interest here was in four neutral problems taken from Raoelison et al. (2020). These neutral problems are basic arithmetic word problems which, unlike conflict or no-conflict problems, are not expected to cue a strong heuristic answer. For example:

In a bar there are forks and knives.
There are 20 forks and twice as many knives.
How many forks and knives are there in total?

These relatively simple problems are traditionally used to track people's knowledge of underlying logico-mathematical building blocks or "mindware" (Stanovich, 2011). Critically, however, although solving the problems requires using similar basic mathematical operations (i.e., addition, multiplication) they do not feature the exact same substitution equation as the bat-and-ball problem (e.g., Y = 2X. X = 20. Y + X =? vs X + Y = 220. Y = X + 200. X =?). Hence, we reasoned that these problems could also be used to test for a potential transfer effect. They allowed us to explore whether the training boosted participant's basic arithmetic word problem solving more generally.

Study 2: Base-rate items. In Study 2, we presented two transfer-neutral base-rate problems from Raoelison et al. (2021), originally created by Pennycook et al. (2014). These problems were specifically designed to avoid any stereotypical associations, as the descriptive information was non-diagnostic. Here is an example of a neutral base-rate problem:

This study contains boys and girls.
Person 'T' is young.
There are 4 boys and 996 girls.
Is Person 'T' more likely to be:
o *a boy*
o *a girl*

In line with previous findings, we also observed in both Study 1 and Study 2 that there was no improvement on the untrained neutral problems (see Supplementary Material Section C for details).

Two-response format

For both Study 1 and Study 2, in both the pre- and post-intervention block, participants responded to each problem using a two-response format, where they first gave a 'fast' answer, immediately followed by a second 'slow' answer (Thompson et al., 2011). This method allowed us to capture both an initial, 'intuitive' response and a final, 'deliberate' one. To minimize the potential for deliberation in the initial response, participants had to provide their first answer within a strict time limit while simultaneously performing a concurrent cognitive load task (see Bago & De Neys, 2017, 2019; Raoelison & De Neys, 2019). The cognitive load task was adapted from the dot memorization task (Miyake et al., 2001), which has been effectively used to burden executive resources in reasoning studies (e.g., De Neys, 2006; Franssens & De Neys, 2009). Participants were required to memorize a complex visual pattern (four crosses in a 3x3 grid) briefly displayed before each reasoning problem. After providing their initial, intuitive response, participants were shown four patterns and had to identify the one they had memorized (see Bago & De Neys, 2019, for more details).

Based on prior pretesting, time limits were set to 5 seconds for bat-and-ball items in Study 1 and 3 seconds for base-rate items in Study 2. These time limits, established in adult samples, reliably induced time pressure and elicited faster responses than traditional one-response formats (Bago & De Neys, 2017, 2019; Boissin et al., 2021, 2022). The simultaneous secondary task further reduced the likelihood of deliberation during the intuitive phase. For the second, final response, participants were allowed to deliberate without any time constraints.

Two-response format and development

The two-response paradigm can be challenging, especially for younger participants. For seventh and eighth graders, the stringent deadline and cognitive load may make reading and understanding the problem more difficult. However, a previous pilot study by Raoelison et al. (2021) showed that the two-response procedure is feasible for early adolescents across various reasoning tasks. Despite the challenges posed by the time limit and load memorization task, most trials were completed successfully. In Study 1, using bat-and-ball problems, early adolescents missed the response deadline on only 15.39% of trials and failed the load task on 16.00% of the

remaining trials. In Study 2, using base-rate problems, 15.52% of trials exceeded the time limit and 15.71% included incorrect responses in the load task. These results confirm that the two-response paradigm is manageable for early adolescents in both tasks.

To further ensure task comprehension among the early adolescents, we analyzed performance on the no-conflict problems presented before the intervention in both studies. These problems are designed such that relying on intuitive reasoning yields the correct answer. If participants could properly read and interpret the problems under the two-response paradigm, they should demonstrate high accuracy. Results confirmed this, with performance on no-conflict trials showing near-ceiling accuracy in Study 1 (initial accuracy: M = 93.44%, SEM = 1.21%; final accuracy: M = 96.88%, SEM = 0.87), and in Study 2 (initial accuracy: M = 83.33%, SEM = 1.66%; final accuracy: M = 79.89%, SEM = 1.93%). These findings indicate that early adolescent participants could indeed read and process the problem material accurately despite the added constraints of the two-response format.

Intervention block

During the intervention block, the participants tried to solve two additional conflict problems without any cognitive or time constraint. In the training group, participants were given an explanation of the correct solution after having responded to each problem. Participants in the control group received no such explanation.

The explanations in each study were translated from English to French based on those used in Boissin et al. (2021, 2022). They were as brief and simple as possible in order to prevent fatigue or disengagement from the task. Importantly, the explanation included both the correct answer and the common incorrect answer. Following Boissin et al. (2021), we also refrained from using personal negative feedback (e.g., 'your answer was incorrect') to reduce any sense of judgment (Trouche et al., 2014). Additionally, to prevent potential mathematical anxiety, the explanation avoided formal algebraic equations (Hoover & Healy, 2017). Full details are presented below.

Study 1: Bat-and-ball items. In Study 1, the intervention block consisted of the standard bat-and-ball problem and one modified version, presented in a fixed order. Unlike the pre- and post-intervention blocks, these bat-and-ball problems used prices instead of quantities. All participants were first presented with the original bat-and-ball problem from (Frederick, 2005). For this problem, participants were asked (1) whether they had encountered the problem before, and (2)

to type their answer and press 'Enter' when ready. They were allowed unlimited time to respond. After responding to each problem, participants—except those in the control group—were shown an explanation of the correct solution as presented in the Supplementary Material Section D.

After receiving an explanation of the previous items (training group) or providing their responses (control group), participants were presented with a second version of the bat-and-ball problem, structurally identical to the standard version but with different surface content (Bago & De Neys, 2019):

A banana and an apple cost $1.40.

The banana costs $1.00 more than the apple. How much does the apple cost?

Study 2: Base-rate items. In Study 2, the intervention block consisted of two base-rate items. After responding to each problem, participants—except those in the control group—were shown an explanation of the correct solution (see Supplementary Material Section D).

Procedure

Debiasing training for bat-and-ball and base-rate tasks followed the same procedure. Each experiment was run individually on computers or tablets, in participants' classrooms, with testing performed in groups under the supervision of at least one teacher and one experimenter. Participants were informed that the experiment would last approximately twenty minutes and required their full attention. They were given a general overview of the task, explaining that they would need to solve reasoning problems and provide two consecutive responses for each problem. Participants were instructed to first give their immediate, initial answer—the response that came to mind first. Following this, they would have the opportunity to reflect and take as much time as needed to provide a final, considered answer. To ensure familiarity with the two-response procedure, participants completed two unrelated practice problems. They then practiced the cognitive load task separately on two load trials, and finally, completed two trials that combined both the cognitive load and two-response procedures.

Figure 1 illustrates a typical trial. To signal that time was running out, the screen background turned yellow 2000 ms before the deadline in the initial stage. If participants failed to respond in time, they were reminded on subsequent trials of the importance of responding within the time limit.

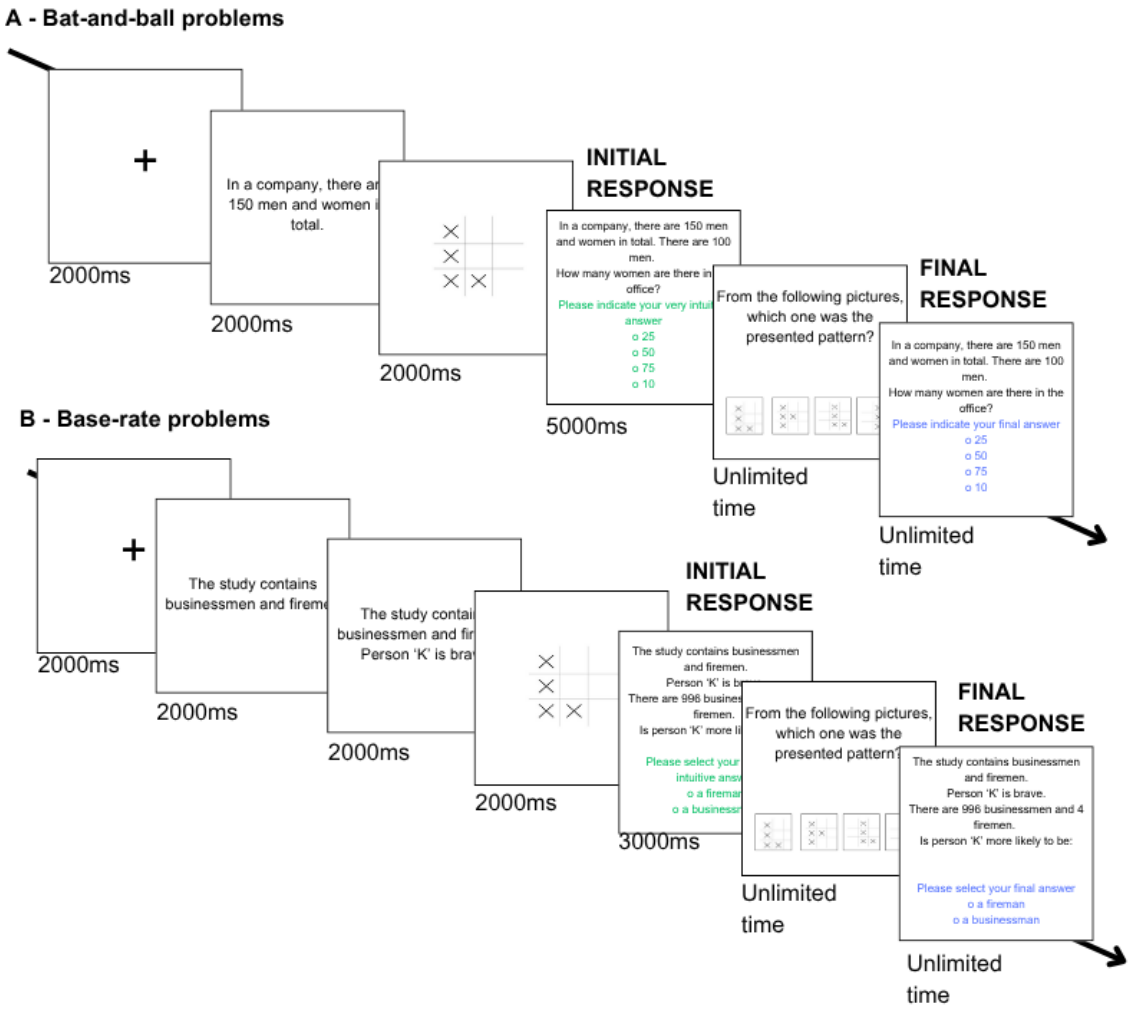

Figure 1. Time course of a complete two-response trial for bat-and-ball items from Study 1 (A) and base-rate items from Study 2 (B).

Trial exclusion

We excluded trials where participants failed to provide their initial response within the allotted time (Study 1: 15.39% for early adolescents, 14.72% for late adolescents; Study 2: 15.52% for early adolescents, 7.64% for late adolescents) or failed to correctly identify the memorized matrix in the load task (Study 1: 16.00% of remaining trials for early adolescents, 14.80% for late adolescents; Study 2: 15.71% for early adolescents, 10.67% for late adolescents). As a result, we analyzed 71.08% of all Study 1 trials for early adolescents, 72.66% for late adolescents and 71.21% of all Study 2 trials for early adolescents and 82.50% for late

adolescents. Note that Fisher's exact tests showed no significant difference between early and late adolescents (all *ps*>.261) in Study 1, however, in Study 2, it shows that early adolescents missed the deadline more often (OR = .46, *p*<.001), failed the memorized matrix more often (OR = .64, *p*<.001) and consequently contributed fewer valid trials (OR = 1.90, *p*<.001). However, the proportion of successful trials remained high (71%), indicating that the paradigm was still feasible for younger participants.

After applying these trial-level exclusion criteria, some participants ended up with no usable response in at least one Block × Response Stage (Initial or Final response stages, in either the Pre- or Post-intervention block). Because computing change scores requires that each participant have at least one response in both blocks for each response stage, these participants could not contribute complete data and were removed from the analyses. In Study 1, this resulted in the exclusion of six early adolescents (three in the control group and three in the training group) and five late adolescents (two in the control group and three in the training group). In Study 2, we excluded only 7 early adolescents (three in the control group and 4 in the training group).

Detailed summaries of the number of valid trials contributed by each age group for each item type are provided in Supplementary Material Section E. Across both studies, early and late adolescents contributed broadly similar amounts of usable data for conflict, no-conflict, and neutral items.

Statistical analyses

The following analyses were conducted in R (R Core Team, 2023) using the bayestestR (Makowski et al., 2019) and brms (Bürkner, 2017) packages. Bayesian model comparisons were performed using Bayes factors (BFs), which indicate how much more likely the observed data are under one model (e.g., including a specific parameter) than under an alternative reduced model (excluding that parameter). We report BF_{01} to denote support for the null hypothesis (i.e., no difference between models indicating that adding the parameter does not increase model support) and BF_{10} to denote support for the alternative hypothesis (i.e., a difference between models indicating greater evidence for the model including the parameter). To ensure clarity, we calculated these factors using matched models, allowing us to identify precisely which hypothesis each BF supports.

We interpret BFs following Jeffreys' (1961) classification scheme, as cited in Lee and Wagenmakers (2014): a BF near 1 indicates no evidence; values between 1 and 3 are considered anecdotal; 3–10, moderate; 10–30, strong; 30–100, very strong; and above 100, extreme.

Therefore, BFs below 3 were interpreted as inconclusive. Traditional inferential analyses are also shown in Supplementary Material Section F.

Results

Conflict items accuracy

For both studies, we tested the effects of training and age using Bayesian matched model comparisons. We first focused on the training effect within each age group by comparing the training and control groups before and after the intervention, separately for initial and final responses. For each response stage, we fit a Bayesian generalized linear mixed-effects model predicting accuracy from Group (Control vs. Training), Block (Pre- vs. Post-intervention), and their interaction (with random intercepts for participants), and compared it to a reduced model without the interaction. Next, the age-moderation comparison model tested whether the training effect differed by age group by comparing a full Bayesian generalized linear mixed-effects model predicting accuracy from Group, Block, Age Group, and their interactions and a random intercept for participants, to a model omitting the three-way interaction. Throughout this section, Bayes Factors index evidence for the interaction term under test: the Group × Block interaction in the training-effect comparison, and the Group × Block × Age Group interaction in the age-moderation comparison.

Study 1: Bat-and-ball items. We first report results from the training-effect comparison model. Figure 2 shows that we replicated patterns previously observed in adult samples (Boissin et al., 2021). That is, for the final responses, adolescents were typically biased before the intervention. After the intervention, late adolescents in the training group showed a clear improvement of 20.52 percentage points (pp, 95% CI [10.99,30.06]), compared with only 4.63pp (95% CI [-4.43, 13.69]) in the control group, BF_{10} = 1050. Among early adolescents, the training group also showed a larger improvement (12.42pp, 95% CI [3.92, 20.93]) than the control group (1.32pp, 95% CI [-1.26, 3.89]), BF_{10} = 15.59.

A similar pattern emerged for initial, intuitive responses. As shown in Figure 2, both early and late adolescents were typically biased before the intervention. After the training, late adolescents in the training group improved by 16.17pp (95% CI [7.68, 24.66]), compared to 5.25pp (95% CI [0.15, 10.34]) in the control group. The training effect was supported by a BF_{10} of 3.32. Among early adolescents, the training group improved by 8.81pp (95% CI [1.58, 16.03]), while the control group improved by only 1.17pp (95% CI [-1.49, 3.83]). The training effect of early adolescents was supported by a BF_{10} of 8.58.

Taken together, both early and late adolescents showed clear improvements after the training for both final and initial responses. A key question, however, is whether this effect is similar across age groups. We then report results from the age-moderation comparison model. For the final responses, the improvement difference between training and control groups was larger for late adolescents (15.83pp) than for early adolescents (10.92pp), indicating a stronger training effect in the older group, BF_{10} = 4.74. For initial responses, the same pattern emerged: late adolescents showed a larger improvement (11.10pp) than early adolescents (7.64pp), BF_{10} = 5.44.

Study 2: Base-rate items. We again begin with the training-effect comparison model. Figure 2 shows that Study 2 revealed a qualitatively similar pattern to Study 1. For final responses, the late adolescent training group improved by 18.71pp (95% CI [10.40, 27.02]), whereas the control group showed a decline of 3.27pp (95% CI [-11.98, 5.45]), BF_{10} = 2890. In the early adolescent group, the training group improved by 16.82pp (95% CI [4.81, 28.84]), while the control group improved by 3.72pp (95% CI [-6.28, 13.72]), BF_{10} = 30.26.

For initial responses, late adolescents in the training group improved by 22.64pp (95% CI [11.27, 34.01]), while the control group showed stable performance, with a change of only 0.65pp (95% CI [-10.32, 11.63]), BF_{10} = 290.45. In the early adolescent group, the training group improved by 16.04pp (95% CI [4.93, 27.14]), and the control group showed again stable performance (95% CI [4.93, 27.14]), BF_{10} = 11.48.

We then report results from the age-moderation comparison model. For final responses, although the training effect was descriptively larger for late adolescents (21.98pp vs. 13.10pp), the age-moderation comparison model yielded extreme evidence against a differential effect across age groups, BF_{01} > 100, suggesting that, contrary to Study 1, both early and late adolescents benefited equally from the intervention in this task, at least for the final deliberate responses.

For initial responses, the difference was again larger for late adolescents (21.99pp vs. 16.04pp for the early adolescents), but the evidence for an age difference was only anecdotal, BF_{10} = 1.82.

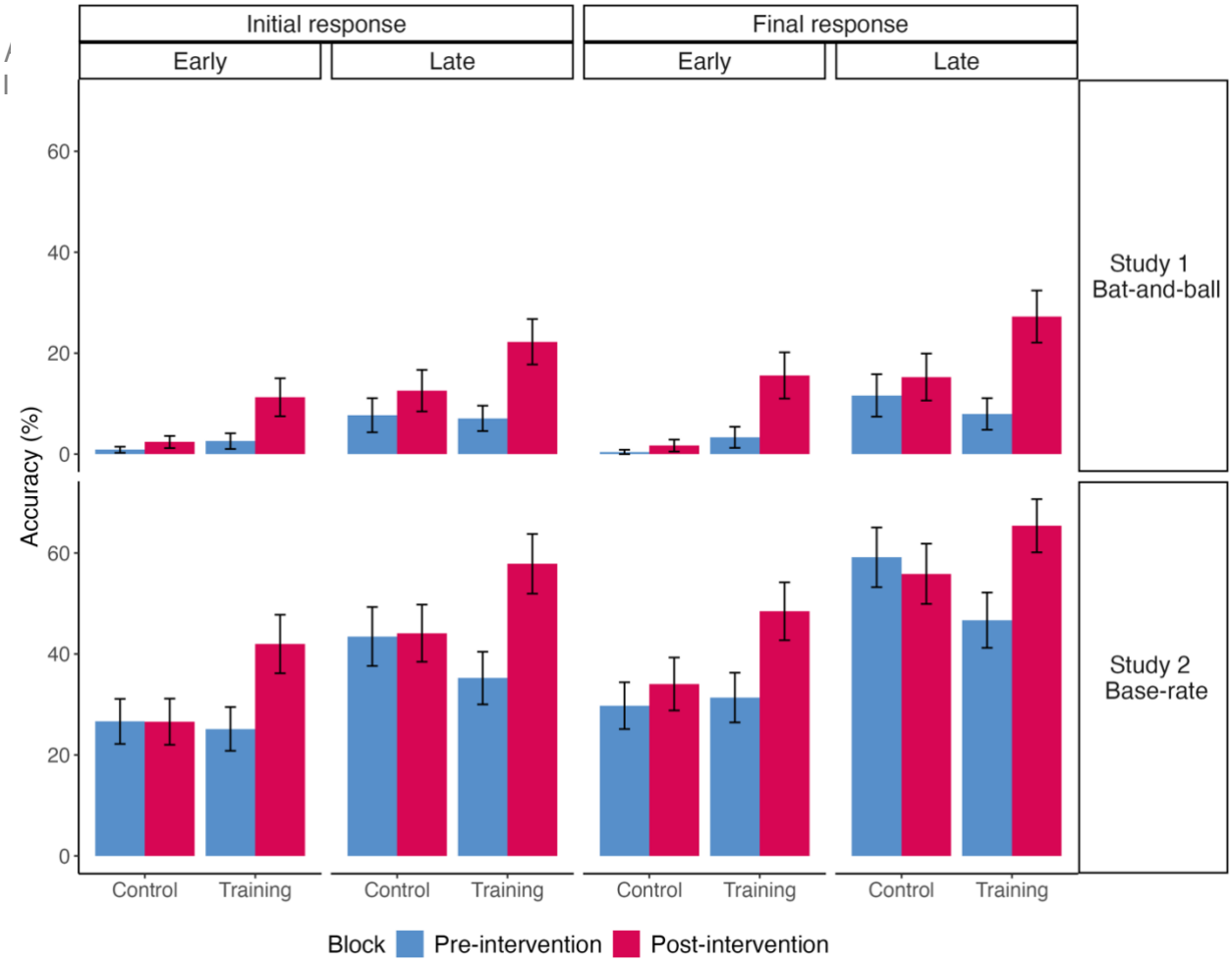

Figure 2. Average initial and final accuracy on conflict problems between early and late adolescents, before and after the intervention for the control and training groups, in Study 1 on bat-and-ball items and Study 2 on base-rate items. Error bars represent standard error of the mean (SEM).

No-conflict items accuracy

As expected, performance on the no-conflict control problems was at ceiling in both studies, for both age groups. In the early adolescent group, accuracy was high in Study 1 (initial: *M* = 93.44%, *SEM* = 1.21%; final: *M* = 96.88%, *SEM* = 0.87%) and in Study 2 (initial: *M* = 83.33%, *SEM* = 1.66%; final: *M* = 79.89%, *SEM* = 1.93%). The same was true for the late adolescent group, with near-ceiling performance in Study 1 (initial: *M* = 94.60%, *SEM* = 1.10%; final: *M* = 96.00%, *SEM* = 1.02%) and high accuracy in Study 2 (initial: *M* = 89.68%, *SEM* = 1.30%; final: *M* = 91.99%, *SEM* = 1.23%). This high level of accuracy indicated that participants successfully processed the problems and were not responding at random. If the initial response constraints had hindered participants' ability to read and understand the problems—particularly in the

younger age group—we would expect to see more random responses on these control no-conflict problems.

Direction of change

To better understand how adolescents corrected (or did not correct) their responses after deliberation, we performed a direction of change analysis on the conflict items (Bago & De Neys, 2017). Especially, 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 different 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.

The central logic is that comparing the change in the proportions of these patterns from pre- to post-intervention allows us to determine which type of improvement (intuitive or deliberative) dominates after the training. A decrease in "00" patterns (i.e., fewer biased responses) combined with a larger increase in "11" than in "01" suggests that training primarily boosts intuitive responding. In contrast, a relatively larger increase in "01" indicates that, even after training, participants still rely on deliberation to correct their initial responses, meaning that the training effect operates mainly through improved deliberative correction. Conversely, persistent "00" patterns signal a lack of improvement. Finally, by comparing these pre–post shifts across early and late adolescents, we assess whether the dominant pattern of improvement differs by age group.

We conducted three matched model comparisons. First, the pattern-shift comparison model tested whether changes in the proportions of "00," "01," and "11" from pre- to post-intervention differed between training and control groups within each age group. We fit a Bayesian linear mixed-effects model predicting proportions from Direction ("00," "01," "11"), Block (Pre vs. Post), Group (Training vs. Control), all their two- and three-way interactions (with random intercepts for participants), and compared to a random intercept for participants to a reduced model omitting the three way interaction.

Next, the within-training comparison model tested whether improvements in the training group were primarily driven by an increase in "11" rather than "01" within each age group. We restricted this analysis to the training group and compared a Bayesian linear mixed-effects model including Direction, Block, their interaction and a random intercept per participant predicting the

proportion of "01" and "11" responses from pre- to post-intervention against a main-effects-only model.

Finally, the age-difference comparison model tested whether the relative increase in "11" over "01" differed between early and late adolescents in the training group, thus, we ran a Bayesian linear mixed-effects model including Direction, Age Group, Block and all 2- and 3-way interactions, with a random intercept per participant, and compared the full model to one without the three-way interaction.

Study 1: Bat-and-ball items. We first report results from the pattern-shift comparison model. Figure 3 shows the direction-of-change distributions for conflict problems before and after the intervention in both control and training groups of early and late adolescents. Late adolescents in the control group maintained relatively stable performance throughout, showing little spontaneous improvement. They consistently produced a majority of "00" responses both before and after the intervention, "01" responses decreased by 2.08pp, and "11" responses increased modestly by 6.55pp. Late adolescents in the training group showed a similar pattern before the intervention, predominantly giving "00" responses, but exhibited a marked shift after the training: "00" responses decreased by 15pp, while "01" and "11" responses increased by 3.81pp and 15.83pp, respectively. The BF_{10} of 32,600, indicates a clear benefit of the training for these participants. Notably, this improvement was primarily driven by an increase in intuitive correct responses ("11"), rather than deliberate corrections ("01"). We fully tested this pattern using the within-training comparison model, BF_{10} = 182.46. This suggests that the training boosted primarily intuitive logical responding among late adolescents, consistent with previous findings in adults (Boissin et al., 2021).

A similar pattern was observed among early adolescents. In the control group, "00" responses remained very high both before and after the intervention, with minimal change in "01" and "11" responses. As with late adolescents, there was little evidence of spontaneous improvement in the control group. In contrast, early adolescents in the training group showed a shift after training: "00" responses decreased by 13.54pp, while "01" responses increased by 3.42pp and "11" responses increased by 8.34pp. This shift was strongly supported by the pattern-shift comparison model, BF_{10} = 70,600.00 again providing extreme evidence for a substantial improvement of participants in the training group. As with late adolescents, this improvement was mainly driven by an increase in intuitive correct responses ("11") rather than deliberate corrections ("01"). This pattern was confirmed by the within-training comparison model, BF_{10} = 21.32.

Finally, according to the age-difference comparison model, the BF_{10} of 29.34, indicated strong evidence for a difference in the pattern of change across age groups, suggesting that the intuitive boost following training was stronger in late than in early adolescents.

Study 2: Base-rate items. We again first report results from the pattern-shift comparison model. Figure 3 shows that the response patterns in Study 2 closely mirrored those observed in Study 1. In the control groups, both early and late adolescents maintained stable proportions of biased "00" responses from pre- to post-intervention, with minimal changes in "01" and "11" responses—again indicating no spontaneous improvement.

In contrast, both training groups showed substantial reductions in "00" responses following the intervention (−20.39pp for early adolescents, −19.97pp for late adolescents). Among early adolescents, this decrease was accompanied by increases in both "01" (+3.83pp) and "11" responses (+13.10pp), BF_{10} = 23,200.00. Notably, the improvement in the training group of early adolescents was mainly driven by an increase in "11" responses, as confirmed by strong support for a differential shift favoring correct intuitions over deliberations in the within-training comparison model, BF_{10} = 47.43.

For late adolescents, the drop in "00" responses was almost entirely accounted for by a substantial increase in "11" responses (+21.38pp), while "01" responses slightly decreased (−2.67pp), as indicated by the pattern-shift comparison among late adolescents; BF_{10} = 119,000.00. The predominance of "11" gains was confirmed by substantial evidence from the within-training comparison model, BF_{10} = 678.56, indicating that the debiasing primarily occurred at the intuitive stage for late adolescents.

Finally, and in line with Study 1, the age-difference comparison model focusing on "01" and "11" responses across training groups revealed that the gain in intuitive accuracy was greater for late than for early adolescents, BF_{10} = 57.91.

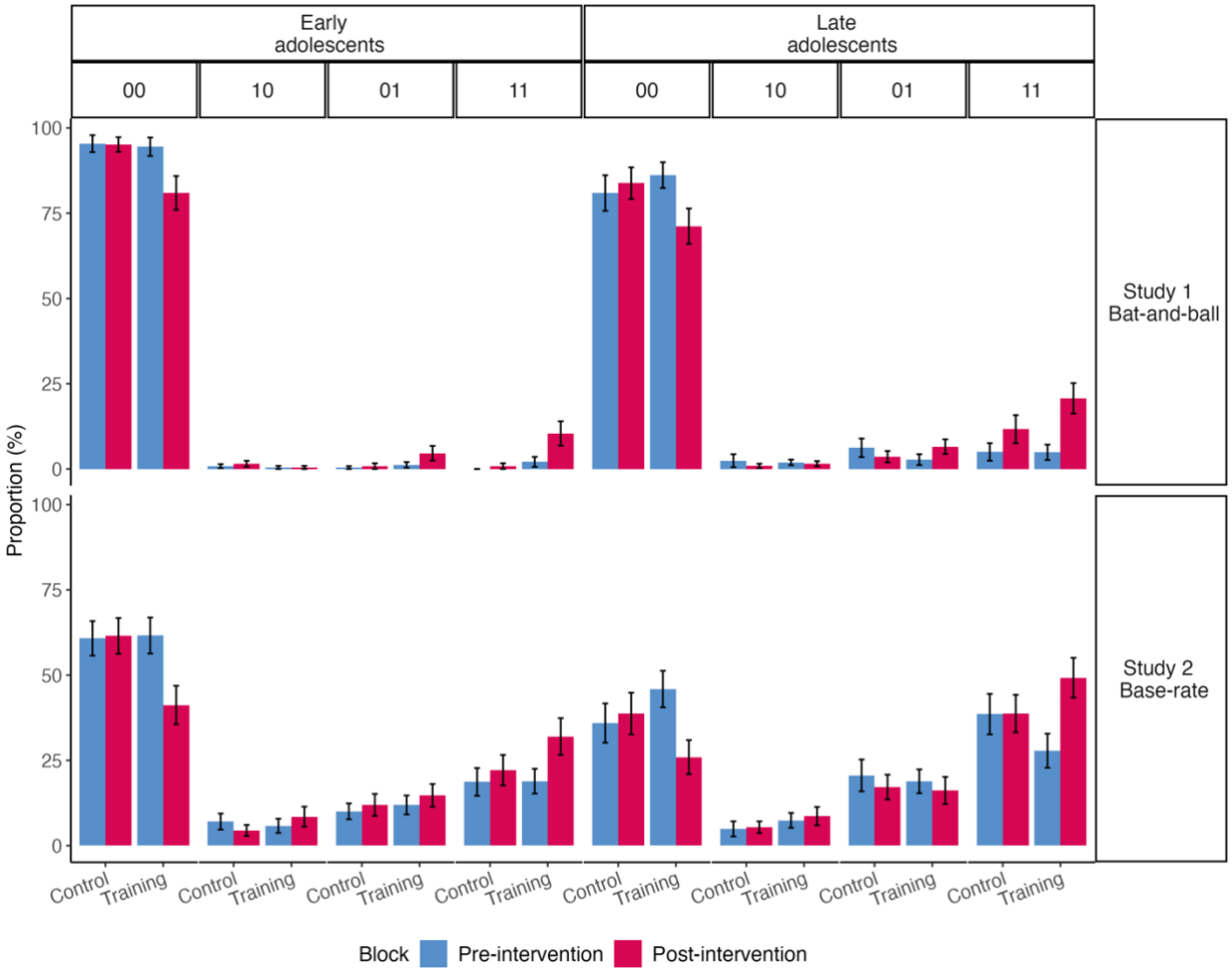

Figure 3. Proportion of each direction of change (i.e., 00 response patterns, 01 response patterns, 10 response patterns and 11 response patterns) for the conflict problems as a function of block, group and age group in Study 1 (bat-and-ball problems) and 2 (base-rate problems). Error bars represent standard error of the mean (SEM).

Discussion

Previous research has shown that a brief, one-time training session explaining the reasoning behind the correct solution can help adults overcome biases and enhance intuitive responding (Boissin et al., 2021, 2022, 2023; Franiatte et al., 2024a, 2024b). This intuitive training effect is thought to rely on the reactivation of prior logical knowledge acquired through formal education. Building on this idea, we hypothesized that late adolescents—who have received more extensive instruction in formal reasoning—would benefit more from the training than their younger

peers. We tested this hypothesis by comparing the training effect in late (11th–12th grade) and early (7th–8th grade) adolescents across two reasoning tasks: the bat-and-ball task (Study 1) and the base-rate task (Study 2), using the two-response paradigm.

In the current study, we used participants' grade level as a proxy for their exposure to logico-mathematical principles, which are progressively introduced and taught throughout secondary education (Hoover & Healy, 2017). As these principles are repeatedly practiced, they become increasingly instantiated and automatized. Consequently, late adolescents are expected to possess a more developed store of automatized logical knowledge than early adolescents (Raoelison et al., 2020). If the training effect relies on reactivating this knowledge (Boissin et al., 2021), its impact should scale with prior exposure: when principles are already well-instantiated, brief explanations should efficiently reactivate them and therefore boost their automatic application.

Following this rationale, one would expect early adolescents—who have had less exposure and thus fewer opportunities to automatize the relevant principles—would benefit less from the intervention, particularly at the intuitive stage. This is exactly what we observed: although early adolescents did show improvement, their gains were smaller than those of their older peers. Nonetheless, the fact that they benefited at all at the intuitive stage suggests that the intervention can still be effective when a certain degree of automatization is already in place.

One may note that although late adolescents showed a stronger intuitive training effect than early adolescents, the difference remained relatively modest: 6.5pp for the bat-and-ball task and 5.8pp for the base-rate task. This small gap may seem surprising given the theoretical role of prior exposure, but there are at least two reasons why the small gap should be interpreted with caution.

First, we used formal education level as a proxy for prior exposure because the core logico-mathematical principles targeted in our tasks are explicitly taught throughout the secondary school curriculum. However, formal instruction is not the only route through which such principles can be acquired. Boissin et al. (2024a) showed that adults from more industrialized Namibian regions benefited more from debiasing training than those from less industrialized rural areas—despite comparable levels of formal education. This suggests that informal education also contributes meaningfully to the development of reasoning and the effectiveness of training. In that sense, early adolescents in our study may already have acquired partial logico-mathematical knowledge through informal contexts, despite their more limited formal instruction.

Second, we compared two age groups within a comparable educational, school context—7th–8th vs. 11th–12th grade. All participants were tested in secondary school to minimize

confounds. However, this also narrowed the gap in practiced knowledge, likely underestimating the full developmental trajectory. Since the automatization of logical intuitions may continue beyond secondary education, larger developmental differences would be expected in comparisons with adults, who have had more time to consolidate logico-mathematical principles. Indeed, adult samples show much larger intuitive training effects. In Boissin et al. (2021, 2022), the proportion of "11" responses increased by 45.8pp on the bat-and-ball task and 48pp on the base-rate task, compared to 15.8pp and 21.4pp among late adolescents in our study. These differences suggest that while late adolescents have begun to instantiate the relevant logical principles, the automatization process may still be incomplete.

In the same vein, if stronger effects emerge with adults, one may also wonder whether more pronounced differences would appear at the other end of the developmental spectrum. For example, one could test younger participants—below 7th grade—to examine whether larger developmental differences emerge when comparing late adolescents to children with even less exposure to the relevant principles. If intuitive training effects scale with prior knowledge, younger children should, in principle, benefit less from the intervention—even less than early adolescents. However, applying the two-response paradigm to younger populations poses clear methodological challenges. The combination of time pressure and cognitive load makes the task less feasible in early primary school, potentially confounding performance with task demands.

A milder version of this issue also appeared within our own sample. Although early adolescents were able to complete the task and showed high accuracy on no-conflict items, the two-response paradigm was nonetheless more demanding for them. In Study 2 in particular, they failed the cognitive load task at a higher rate than late adolescents. This pattern does not challenge the main findings, but it indicates that task demands were not entirely equivalent across age groups. These considerations slightly qualify the developmental interpretation by suggesting that the observed age difference may reflect both differences in prior knowledge and differences in how younger participants managed the task under time pressure and cognitive load. Importantly, this nuance does not alter our conclusions: younger adolescents still produced valid trials at a high rate (71%), indicating that the paradigm remained feasible and that the age contrast is not reducible to task failure.

To avoid confusion, it is important to note that our results do not imply that it is impossible to debias early adolescents' intuitions or that they are necessarily less responsive to such interventions. Clearly, with the current intervention's design, we focused on reasoning tasks and logico-mathematical principles with which earlier adolescents are less familiar and received less instruction. Nevertheless, we believe that these same adolescents could still benefit from

debiasing interventions that activate previously acquired knowledge. For instance, future studies could investigate the impact of training on arithmetic problems, such as those typically taught in primary school (e.g., "Mary has 8 marbles. She has 5 more marbles than John. How many marbles does John have?", Riley et al., 1983, cited in Lubin et al., 2015), or on the comparison of decimal numbers (e.g., the erroneous belief among school-aged children that 1.45 is greater than 1.5 because 45 is greater than 5; Wearne & Hiebert, 1988). Since these principles are taught and practiced at younger (elementary school) ages, they may speculatively show much stronger intuitive debiasing effects.

On a more methodological perspective, we also note that accuracy and direction-of-change analyses can lead to different conclusions—a distinction that matters for interpreting the effect of the training or more generally, the dynamic correction—or non-correction—of intuitive responses. Specifically, on the base-rate task, both early and late adolescents improved following the training, and accuracy analyses showed comparable gains at both the intuitive and final response stages. Based on these results, one might conclude that the training was equally effective across age groups. However, the direction-of-change analysis revealed a different picture. Although both groups showed an increase in “11” responses—indicating correct intuitive responding—the gains were larger among late adolescents, suggesting a stronger intuitive boost in this group. Unlike accuracy, which captures response outcomes but not their temporal dynamics, direction-of-change tracks how answers evolve from the initial to the final stage. It allows us to distinguish responses that were already correct at the intuitive stage from those corrected through deliberation. This distinction is essential: relying solely on final accuracy makes it impossible to conclude whether a correct answer was generated intuitively or arrived at through later correction. As a result, improvements may be wrongly attributed to better deliberation when they in fact reflect improved intuitions. In our study, most final correct responses were already correct at the initial stage—indicating that the training primarily boosted intuitive performance. Direction-of-change analysis captured this subtlety by showing that the improvement lay not only in the number of correct responses, but in how those responses were generated—through intuitive rather than deliberative reasoning. When the goal is to assess the nature of sound reasoning, this method offers a more precise and informative perspective.

From a developmental perspective, the present findings are also compatible with broader accounts of how intuitive reasoning improves with age and experience. For example, fuzzy-trace theory proposes that development involves a shift toward more abstract and task-relevant representations, allowing reasoners to respond rapidly without reconstructing all details of a problem (Reyna, 2012; Reyna & Brainerd, 2011). Although this framework does not explicitly

invoke automatization, it converges with our interpretation in a key respect: with learning and schooling, correct responses become easier to generate intuitively because task-diagnostic relations can be accessed more efficiently.

At the same time, an important question remains open: what exactly does the training reactivate? While our data show that prior learning predicts the extent of intuitive debiasing, they do not identify the nature of the underlying change. One possibility is that training facilitates the direct retrieval of well-learned response patterns or logico-mathematical principles (e.g., "the larger group matters more"), allowing participants to respond correctly without further computation. Another possibility is that training helps participants construct a more appropriate understanding of the problem structure (e.g., recognizing which information is relevant and which can be ignored), making the correct response easier to derive even without a stored procedure (Markovits et al., 2019; Markovits & Barrouillet, 2004). The present data do not allow us to distinguish between these possibilities, leaving open what is reactivated by training.

It is important to highlight that the logical nature of the intuitions remains debated in the literature. Critics argue that correct intuitive responses do not necessarily reflect genuine sensitivity to formal structure (Ghasemi, Handley, Howarth, et al., 2022; Ghasemi, Handley, & Stephens, 2022; Handley et al., 2023; Hayes et al., 2022; Mekik et al., 2025; Meyer-Grant et al., 2022). Instead, such responses may be driven by superficial heuristic cues that merely happen to coincide with the logical correct answer. According to this view, there is no intrinsic or epistemic link between intuitive correctness and logic. What is often labelled a "logical intuition" may instead reflect the activation of domain-specific heuristics that mimic formal reasoning without engaging with its underlying structure. In this sense, intuitive logical reasoning would serve to calculate a proxy of logical reasoning but not actual logical reasoning. That is, against the present developmental backdrop, one may also argue that schooling through secondary education may also be critical to develop the alleged heuristic proxies. Nonetheless, this theoretical debate does not undermine the empirical finding that late adolescents benefit more from training and are able to produce correct responses intuitively. Their performance suggests that, regardless of whether the underlying mechanism is genuinely logical or heuristic in nature, prior exposure facilitates the generation of correct responses without the need for further deliberation.

More broadly, the present findings align with a large body of work on the development of numerical and mathematical skills. In these domains, learning is typically not driven by the discovery of new principles, but by the gradual stabilization and routinization of ways of producing correct answers through repeated instruction and practice (e.g., Rittle-Johnson & Siegler, 2001; Siegler, 1996; Woodward, 2006). Early in learning, correct responding often requires explicit,

effortful steps; with schooling, these steps become increasingly compressed, making correct responses easier to generate and less dependent on deliberation.

Applied to the present context, this perspective suggests that classic reasoning problems may follow a similar developmental trajectory. Brief explanations are sufficient to shift responding when relevant principles have already been partially instantiated, but not when they are largely absent. Accordingly, the stronger intuitive gains observed among late adolescents are better understood as reflecting a more advanced stage of consolidation rather than a qualitative difference in reasoning mechanisms. In this sense, our contribution is to show that a learning dynamic well documented in mathematics and numeracy can also be observed in classic reasoning tasks, provided that prior knowledge has reached a sufficient level of consolidation.

Taken together, our findings show that intuitive debiasing is possible in adolescents, but its effectiveness depends not only on the structure of the intervention, but also on what the learner already knows. Training is most effective when it reactivates principles that are already, at least partially, instantiated. This supports the idea that logical intuitions can emerge from prior learning and be strengthened through targeted explanation—especially when learners have begun to instantiate the relevant principles. Ultimately, the success of intuitive debiasing relies on a simple but critical insight: instruction is most powerful when it builds on knowledge that is already in place.

Fundings details

This research was supported by a grant from the Agence Nationale de la Recherche (ANR-23-CE28-0004-01) and by The Fyssen Foundation.

Disclosure statement

The authors report there are no competing interests to declare.

Data availability statement

The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/96avw/?view_only=f0ae615c9ac04c1c899d3fab640129dd.

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 & Reasoning*, *25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Boissin, E., Caparos, S., & De Neys, W. (2023). Examining the role of deliberation in de-bias training. *Thinking & Reasoning*, *0*(0), 1–29. https://doi.org/10.1080/13546783.2023.2259542

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

Boissin, E., Caparos, S., Voudouri, A., & De Neys, W. (2022). Debiasing System 1: Training favours logical over stereotypical intuiting. *Judgment and Decision Making*, *17*(4), 646–690. https://doi.org/10.1017/S1930297500008895

Bourgeois-Gironde, S., & Van der Henst, J.-B. (2009). How to open the door to System 2: Debiasing the Bat-and-Ball problem. *Rational Animals, Irrational Humans*, 235–252.

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

Claidière, N., Trouche, E., & Mercier, H. (2017). Argumentation and the diffusion of counter-intuitive beliefs. *Journal of Experimental Psychology: General*, *146*(7), 1052–1066. https://doi.org/10.1037/xge0000323

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. (2023). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, *46*, e111. 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

Evans, J. St. 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

Franiatte, N., Boissin, E., Delmas, A., & Neys, W. D. (2024). Adieu Bias: Debiasing Intuitions Among French Speakers. *Psychologica Belgica*, *64*(1), 42. https://doi.org/10.5334/pb.1260

Franiatte, N., Boissin, E., Delmas, A., & De Neys, W. (2024). Boosting debiasing: Impact of repeated training on reasoning. *Learning and Instruction*, *89*, 101845. https://doi.org/10.1016/j.learninstruc.2023.101845

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

Ghasemi, O., Handley, S., Howarth, S., Newman, I. R., & Thompson, V. A. (2022). Logical intuition is not really about logic. *Journal of Experimental Psychology: General*, *151*, 2009–2028. https://doi.org/10.1037/xge0001179

Ghasemi, O., Handley, S., & Stephens, R. (2022). *Logical intuitions or matching heuristic? Examining the effect of deduction training on belief-based reasoning judgments*. OSF. https://doi.org/10.31219/osf.io/gxd73

Handley, S. J., Ghasemi, O., & Bialek, M. (2023). Illusory intuitions: Challenging the claim of non-exclusivity. *The Behavioral and Brain Sciences*, *46*, e125. https://doi.org/10.1017/s0140525x22003168

Hayes, B. K., Stephens, R. G., Lee, M. D., Dunn, J. C., Kaluve, A., Choi-Christou, J., & Cruz, N. (2022). Always look on the bright side of logic? Testing explanations of intuitive sensitivity to logic in perceptual tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(11), 1598–1617. https://doi.org/10.1037/xlm0001105

Hoover, J., & Healy, A. (2021). The bat-and-ball problem: A word-problem debiasing approach. *Thinking & Reasoning*, *27*(4), 567–598. https://doi.org/10.1080/13546783.2021.1878473

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*. Macmillan.

Kahneman, D., & Frederick, S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases* (1st ed., pp. 49–81). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.004

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

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare? *Perspectives on Psychological Science*, *4*(4), 390–398. https://doi.org/10.1111/j.1745-6924.2009.01144.x

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

Mekik, C., Vivier, O., & Markovits, H. (2025). A “logical intuition” based on semantic associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001468

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision Making*, *13*(3), 246–259. https://doi.org/10.1017/S1930297500007683

Meyer-Grant, C. G., Cruz, N., Singmann, H., Winiger, S., Goswami, S., Hayes, B. K., & Klauer, K. C. (2022). Are logical intuitions only make-believe? Reexamining the logic-liking effect. - PsycNET. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*, Advance online publication. https://doi.org/10.1037/xlm0001152

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How Can Decision Making Be Improved? *Perspectives on Psychological Science*, *4*(4), 379–383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. *Trends in Cognitive Sciences*, *14*(10), 435–440. https://doi.org/10.1016/j.tics.2010.07.004

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing Decisions: Improved Decision Making With a Single Training Intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129–140. https://doi.org/10.1177/2372732215600886

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

Nisbett, R. E. (Ed.). (1993). *Rules for Reasoning*. Psychology Press. https://doi.org/10.4324/9780203763230

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 544–554. https://doi.org/10.1037/a0034887

Purcell, Z. A., Wastell, C. A., & Sweller, N. (2021). Domain-specific experience and dual-process thinking. *Thinking & Reasoning*, 1–29. https://doi.org/10.1080/13546783.2020.1793813

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

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

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory. *Judgment and Decision Making*, *7*(3), 332–359. https://doi.org/10.1017/S1930297500002291

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

Rittle-Johnson, B., & Siegler, R. S. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346–362.

Siegler, R. S. (1998). *Emerging minds: The process of change in children's thinking*. Oxford University Press.

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision Making*, *13*(3), 260–267.

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

Trouche, E., Sander, E., & Mercier, H. (2014). Arguments, more than confidence, explain the good performance of reasoning groups. *Journal of Experimental Psychology: General*, *143*(5), 1958–1971. https://doi.org/10.1037/a0037099

Woodward, J. (2006). *Developing automaticity in multiplication facts: Integrating strategy instruction with timed practice drills*. *Journal of Special Education, 40*(3), 151–162.

Supplementary Material

A. Study 1's results without participants knowing the bat-and-ball problems and giving a correct response

Conflict items accuracy

Table A1. Mean percentage of Conflict accuracies (with standard errors of the mean) in the Control and Training groups, for both Early and Late adolescents who either did not know the traditional bat-and-ball item or knew it but answered incorrectly.

	Late				Early			
	Training n=60		Control n=58		Training n=56		Control n=60	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Initial response	4.53 (2.12)	14.44 (4.03)	2.17 (2.17)	5.21 (2.97)	2.58 (1.56)	11.27 (3.76)	0.86 (0.60)	2.40 (1.21)
Final response	3.95 (2.49)	16.81 (4.56)	5.43 (3.10)	6.25 (3.24)	3.33 (2.09)	15.59 (4.58)	0.43 (0.43)	1.69 (1.19)

No-conflict items accuracy

Table A2. Mean percentage of No-conflict accuracies (with standard errors of the mean) in the Control and Training groups, for both Early and Late adolescents who either did not know the traditional bat-and-ball item or knew it but answered incorrectly.

	Late		Early	
	Training n=60	Control n=58	Training n=56	Control n=60

	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Initial response	96.84 (1.49)	96.25 (1.96)	96.53 (1.79)	98.26 (0.99)	98.33 (0.95)	95.68 (1.34)	89.55 (3.23)	90.69 (2.89)
Final response	98.99 (0.71)	95.97 (2.02)	97.05 (1.48)	99.48 (0.52)	99.39 (0.61)	97.02 (1.17)	95.06 (2.35)	96.25 (2.04)

Transfer-neutral items accuracy

Table A3. Mean percentage of Transfer-neutral accuracies (with standard errors of the mean) in the Control and Training groups, for both Early and Late adolescents who either did not know the traditional bat-and-ball item or knew it but answered incorrectly.

	Late				Early			
	Training		Control		Training		Control	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Initial response	50.00 (9.28)	34.78 (6.38)	72.00 (8.21)	32.43 (7.31)	45.31 (8.51)	26.92 (6.05)	45.95 (7.60)	25.64 (6.33)
Final response	87.93 (5.90)	67.39 (6.06)	82.00 (7.57)	64.86 (7.72)	76.56 (7.44)	51.28 (7.46)	74.32 (6.88)	53.85 (7.20)

B. Study 2's pilot rating task

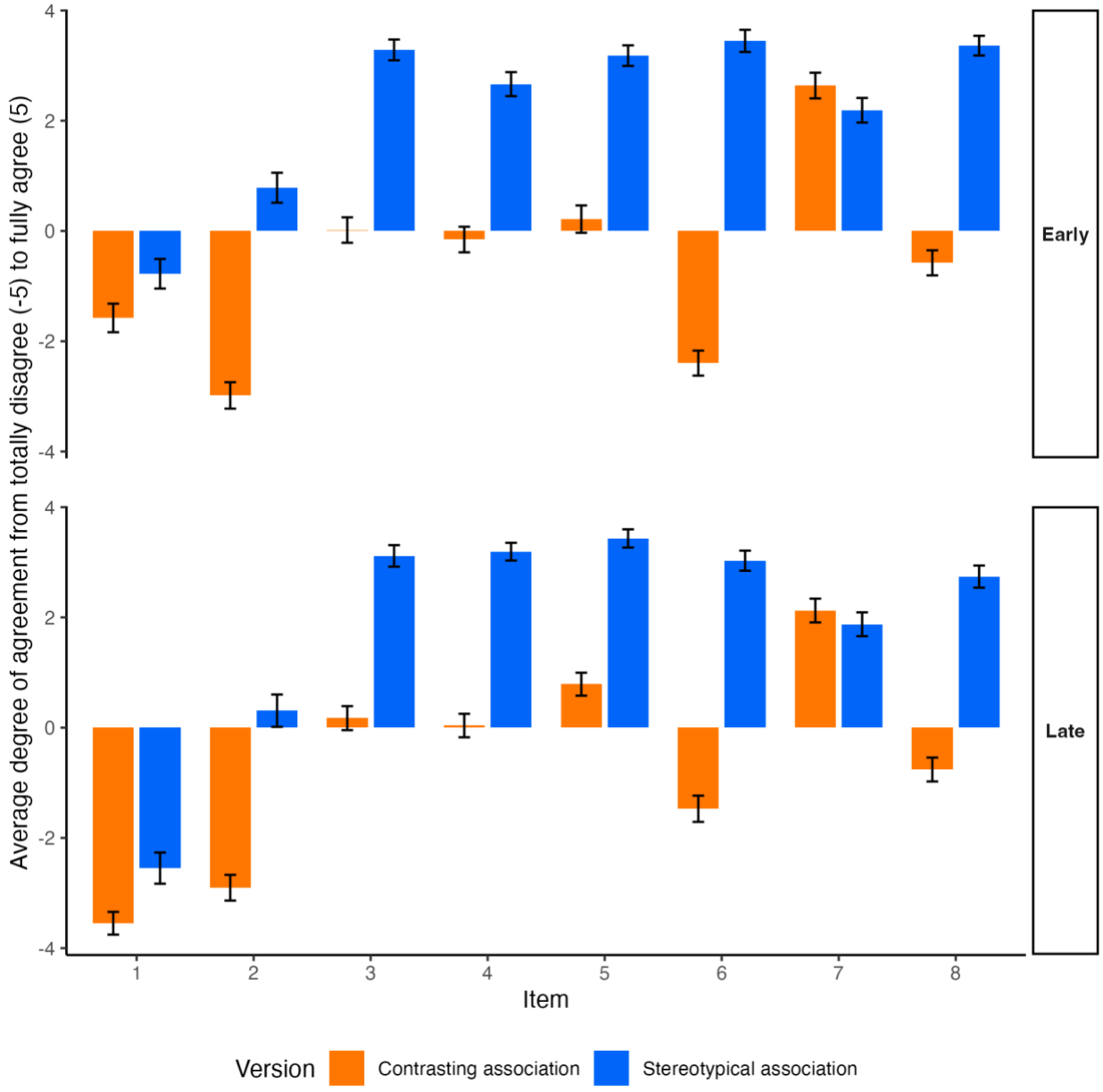

Figure B1. Average degree of agreement of each base-rate trial as a function of the association (stereotypical vs contrasting association). Bar errors are errors standard of the mean (SEM). Validated items are those for which the score of agreement of stereotypical statements is higher than the score of the contrasting association.

C. Transfer-neutral accuracies for Study 1 and Study 2

Table C1. Average initial and final accuracy (standard errors of the mean, SEM) on neutral problems in Study 1 (bat-and-ball problems) and Study 2 (base-rate problems) as a factor of age group (early vs late adolescents) and group (control vs training) before and after the intervention.

			Late		Early	
		Block	Training	Control	Training	Control
Study 1 (bat-and-ball problems)	Initial response	Pre	55.71 (8.39)	67.74 (7.88)	45.31 (8.51)	45.95 (7.60)
		Post	40.91 (6.10)	32.14 (6.78)	26.92 (6.05)	25.64 (6.33)
	Final response	Pre	87.14 (5.55)	85.48 (6.22)	76.56 (7.44)	74.32 (6.88)
		Post	71.82 (5.31)	69.05 (7.02)	51.28 (7.46)	53.85 (7.20)
Study 2 (base-rate problems)	Initial response	Pre	59.38 (6.43)	76.04 (5.37)	41.30 (6.28)	50.00 (5.87)
		Post	71.00 (4.76)	89.13 (3.44)	61.76 (5.16)	54.90 (5.82)
	Final response	Pre	67.71 (6.23)	85.42 (4.46)	60.87 (6.94)	55.66 (5.97)
		Post	85.00 (4.10)	90.22 (3.68)	69.61 (5.44)	52.94 (6.01)

D. Intervention block: Explanation phase

Study 1 - Bat-and-ball items

The correct answer to the previous problem is 5 cents. Many people think it is 10 cents, but this answer is wrong.
If the ball costs 10 cents, the bat would cost $1.10 (as it costs $1.00 more than the ball); both together, they would then cost $1.20.
However, the problem said they cost $1.10 together.
The correct response is that the ball costs 5 cents, the bat $1.05 so together they cost $1.10 ($0.05 + $1.05 = $1.10).

Study 2 - Base-rate items

The correct answer to the previous problem is that person 'K' is most likely a "businessman". Many people think it is "fireman", but this answer is wrong.
Most people base their answer solely on the description ("Person K is brave"). If this were all the information you got, this answer would be correct, as it is likely that there are more brave firemen in the world than brave businessmen.
However, in the problem you also got information about the specific number of businessmen and firemen in the group that person K got drawn from. You were informed that person K was drawn randomly from a group with 996 businessmen and only 4 firemen. Since there are so many more businessmen in the group than firemen (200 times more!), it becomes more likely that person K is a businessman. After all, although firemen might in general be braver than businessmen, there are also some businessmen who are brave. If you combine this with the vastly larger number of businessmen in the group, it will be more plausible that you're dealing with a brave businessman.

E. Average trial contribution

In Study 1, participants in the early adolescent group contributed on average 6.03 (SEM = 0.13) conflict trials out of 8, 6.51 (SEM = 0.13) no-conflict trials out of 8, and 2.21 (SEM = 0.11) transfer-neutral trials out of 4. Late adolescents contributed 6.10 (SEM = 0.12) conflict trials, 6.71 (SEM = 0.12) no-conflict trials, and 2.06 (SEM = 0.09) transfer-neutral trials. In Study 2, early adolescents contributed an average of 5.69 (SEM = 0.16) conflict trials out of 8, 5.96 (SEM = 0.15) no-conflict trials, and 2.98 (SEM = 0.09) transfer-neutral trials out of 4. Late adolescents

contributed 6.69 (SEM = 0.11) conflict trials, 6.62 (SEM = 0.14) no-conflict trials, and 3.23 (SEM = 0.09) transfer-neutral trials.

F. Inferential analyses

1. Conflict accuracies

Early vs Late adolescents

Table E1. Generalized linear mixed-effects model (logit link) predicting accuracy on conflict items across age groups for both initial and final responses, in Study 1 (bat-and-ball items, Firth-penalized logistic regression) and Study 2 (base-rate items). Fixed effects include Age Group, Group, Block, and their interactions. A random intercept was included for Subject.

		Initial response accuracy			Final response accuracy		
	Predictors	Odds Ratio	CI	*p*	Odds Ratio	CI	*p*
Study 1 - Bat-and-ball items	Intercept	0.02	0.00 – 0.05	<.001	0.01	0.00 – 0.05	<.001
	AgeGroup (Late)	5.31	1.32 – 21.30	.019	13.22	2.44 – 71.798	.003
	Group (Training)	1.88	0.39 – 9.00	.430	3.88	0.63 – 24.02	.145
	Block (Post)	1.63	0.34 – 7.81	.539	1.50	0.20 – 11.58	.695
	AgeGroup x Group	0.61	0.11 – 3.52	.581	0.19	0.03 – 1.33	.093

	AgeGroup x Block	1.04	0.18 – 5.89	.965	0.88	0.10 –7.48	.904
	Group x Block	2.67	0.41 – 17.56	.306	3.59	0.38 – 33.94	.265
	AgeGroup x Group x Block	0.72	0.09 – 5.92	.758	0.91	0.08 – 10.20	.939
Study 2 - Base-rate items	Intercept	0.22	0.11 – 0.42	<.001	0.27	0.13 – 0.56	<.001
	AgeGroup (Late)	2.85	1.12 – 7.27	.028	6.74	2.36 – 19.24	<.001
	Group (Training)	0.99	0.39 – 2.53	.985	0.94	0.34 – 2.62	.910
	Block (Post)	0.98	0.53 – 1.81	.943	1.15	0.63 – 2.08	.648
	AgeGroup x Group	0.58	0.15 – 2.20	.425	0.42	0.10 – 1.81	.243
	AgeGroup x Block	1.32	0.58 – 3.02	.512	0.76	0.33 – 1.75	.513
	Group x Block	2.66	1.12 – 6.33	.027	3.18	1.35 – 7.50	.008
	AgeGroup x	1.43	0.44 – 4.66	.558	1.61	0.48 – 5.39	.438

	Group x Block						

Note. Odds ratios and 95% confidence intervals (CI) are reported. p values are based on Wald tests. Reference levels are: AgeGroup = Early, Group = Control, Block = Pre-intervention.

Early adolescents

Table E2. Generalized linear mixed-effects model (logit link) predicting accuracy on conflict items among early adolescents both on initial and final responses, in Study 1 (bat-and-ball items, Firth-penalized logistic regression) and Study 2 (base-rate items). Fixed effects include Group, Block, and their interaction. A random intercept was included for Subject.

		Initial response accuracy			Final response accuracy		
	Predictors	Odds Ratio	CI	p	Odds Ratio	CI	p
Study 1 - Bat-and-ball items	Intercept	0.02	0.00 - 0.05	<.001	0.00	0.01 - 0.05	<.001
	Group (Training)	1.88	0.39 – 9.01	.429	3.88	0.63 – 14.06	.145
	Block (Post)	1.27	0.25 –6.5’	.778	1.51	0.20 – 11.62	.693
	Group x Block	3.44	0.49 – 24.12	.214	3.58	0.38 – 33.86	.266
Study 2 - Base-rate	Intercept	0.24	0.13 – 0.43	<.001	0.31	0.17 – 0.57	<.001

items	Group (Training)	1.01	0.44 – 2.28	.988	0.95	0.41 – 2.22	.909
	Block (Post)	0.96	0.53 – 1.74	.900	1.12	0.63 – 1.98	.698
	Group x Block	2.53	1.10 – 5.83	.029	2.89	1.27 – 6.58	.011

Note. Odds ratios and 95% confidence intervals (CI) are reported. p values are based on Wald tests. Reference levels are: Group = Control, Block = Pre-intervention.

Late adolescents

Table E3. Generalized linear mixed-effects model (logit link) predicting accuracy on conflict items among late adolescents both on initial and final responses, in Study 1 (bat-and-ball items) and Study 2 (base-rate items). Fixed effects include Group, Block, and their interaction. A random intercept was included for Subject.

		Initial response accuracy			Final response accuracy		
	Predictors	Odds Ratio	CI	p	Odds Ratio	CI	p
Study 1 - Bat-and-ball items	Intercept	0.08	0.04 – 0.15	<.001	0.12	0.07 – 0.20	<.001
	Group (Training)	1.15	0.52 – 2.52	.732	0.72	0.35 – 1.50	.379
	Block (Post)	1.76	.83 – 3.72	.138	1.37	0.71 – 2.65	.354

	Group x Block	1.91	0.74 – 4.97	.182	3.26	1.34 – 7.96	.009
Study 2 - Base-rate items	Intercept	0.60	0.28 – 1.27	.183	1.99	0.81 – 4.93	.135
	Group (Training)	0.55	0.19 – 1.60	.275	0.36	0.10 – 1.27	.112
	Block (Post)	1.30	0.74 – 2.31	.365	0.85	0.46 – 1.58	.609
	Group x Block	4.13	1.77 – 9.63	<.001	5.82	2.37 – 14.32	<.001

Note. Odds ratios and 95% confidence intervals (CI) are reported. p values are based on Wald tests. Reference levels are: Group = Control, Block = Pre-intervention.

2. Direction of change

Early adolescents

Table E4. Linear mixed-effects model predicting proportions of “00”, “01” and “11” response patterns among early adolescents as a function of Direction, Block, and Group, including interaction terms. The model included a random intercept for Subject.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	95.42	91.03 – 99.80	<.001
Direction [01]	-95.00	-101.20 – -88.80	<.001
Direction [11]	-95.42	-101.62 – -89.21	<.001
Block [Post-intervention]	-0.28	-6.48 – 5.92	.930
Group [Training group]	-0.92	-7.23 – 5.39	.774

Direction [01] × Block [Post-intervention]	0.69	-8.08 – 9.47	.877
Direction [11] × Block [Post-intervention]	1.11	-7.66 – 9.88	.804
Direction [01] × Group [Training group]	1.70	-7.23 – 10.62	.709
Direction [11] × Group [Training group]	3.01	-5.92 – 11.93	.509
Block [Post-intervention] × Group [Training group]	-13.26	-22.19 – -4.34	.004
(Direction [01] × Block [Post-intervention]) × Group [Training group]	16.27	3.65 – 28.89	.012
(Direction [11] × Block [Post-intervention]) × Group [Training group]	20.76	8.14 – 33.39	.001

Note. 95% confidence intervals (CI) are reported. Reference levels are: Group = Control, Block = Pre-intervention, Direction = '00'.

Table E5. Linear mixed-effects model predicting proportions of "01" and "11" response patterns among early adolescents in the Training group as a function of Direction and Block, including their interaction. The model included a random intercept for Subject.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.19	-3.25 – 5.63	.598
Direction [11]	0.89	-4.82 – 6.61	.758
Block [Post-intervention]	3.42	-2.29 – 9.14	.239
Direction [11] × Block [Post-intervention]	4.91	-3.17 – 12.99	.232

Note. 95% confidence intervals (CI) are reported. Reference levels are: Block = Pre-intervention, Direction = '01'.

Late adolescents

Table E6. Linear mixed-effects model predicting proportions of "00", "01" and "11" response patterns among late adolescents as a function of Direction, Block, and Group, including interaction terms. The model included a random intercept for Subject.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	83.04	75.59 – 90.48	<.001
Direction [01]	-76.93	-87.46 – -66.41	<.001
Direction [11]	-78.12	-88.65 – -67.60	<.001
Block [Post-intervention]	0.60	-9.93 – 11.12	.912
Group [Training group]	3.15	-6.83 – 13.14	.535
Direction [01] × Block [Post-intervention]	-2.68	-17.56 – 12.20	.724
Direction [11] × Block [Post-intervention]	5.95	-8.93 – 20.84	.433
Direction [01] × Group [Training group]	-6.52	-20.64 – 7.60	.365
Direction [11] × Group [Training group]	-3.18	-17.30 – 10.93	.658
Block [Post-intervention] × Group [Training group]	-15.60	-29.71 – -1.48	.030
(Direction [01] × Block [Post-intervention]) × Group [Training group]	21.49	1.52 – 41.46	.035
(Direction [11] × Block [Post-intervention]) × Group [Training group]	24.88	4.91 – 44.85	.015

Note. 95% confidence intervals (CI) are reported. Reference levels are: Group = Control, Block = Pre-intervention, Direction = '00'.

Table E7. Linear mixed-effects model predicting proportions of "01" and "11" response patterns among late adolescents in the Training group as a function of Direction and Block, including their interaction. The model included a random intercept for Subject.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	2.74	-2.84 – 8.31	.334
Direction [11]	2.14	-5.00 – 9.29	.555
Block [Post-intervention]	3.81	-3.33 – 10.95	.295
Direction [11] × Block [Post-intervention]	12.02	1.92 – 22.13	**.020**

Note. 95% confidence intervals (CI) are reported. Reference levels are: Block = Pre-intervention, Direction = '01'.

Early vs Late adolescents

Table E8. Linear mixed-effects model predicting proportions of "01" and "11" response patterns for adolescents in the Training group as a function of Direction, Age Group and Block including interaction terms. The model included a random intercept for Subject.

Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.19	-4.31 – 6.69	.671
Direction [11]	0.89	-6.16 – 7.95	.804
Block [Post-intervention]	3.42	-3.63 – 10.48	.341
AgeGroup [Late adolescents]	1.55	-5.83 – 8.92	.680

Direction [11] × Block [Post-intervention]	4.91	-5.06 – 14.89	.334
Direction [11] × AgeGroup [Late adolescents]	1.25	-8.21 – 10.71	.795
Block [Post-intervention] × AgeGroup [Late adolescents]	0.39	-9.08 – 9.85	.936
(Direction [11] × Block [Post-intervention]) × AgeGroup [Late adolescents]	7.11	-6.27 – 20.50	.297

Note. 95% confidence intervals (CI) are reported. Reference levels are: Group = Control, Block = Pre-intervention, Direction = '00', AgeGroup = Early adolescents..