Trends in Neuroscience and Education 44 (2026) 100295

Contents lists available at ScienceDirect

Trends in Neuroscience and Education

journal homepage: www.elsevier.com/locate/tine

Research paper

Developing system 1: knowledge automatization drives sound intuiting

Esther Boissin [a,*], Laura Charbit [b], Elora Taieb [b], Grégoire Borst [b], Wim De Neys [b,c]

[a] Department of Psychology, Cornell University, Uris Hall 211, 109 Tower Rd, Ithaca, NY, 14853, United States
[b] LaPsyDÉ (UMR 8240), Université Paris Cité, 46 rue Saint-Jacques, 75005, Paris, France
[c] CNRS (Centre national de la recherche scientifique), 3 rue Michel-Ange, 75794 Paris Cedex 16, France

ARTICLE INFO

Keywords:
Automatization
Intuition
Adolescence
Reasoning
Mindware
Numeracy
Dual-process

ABSTRACT

Recent dual-process research shows that adults can generate correct responses intuitively, whereas younger reasoners rarely do so. Yet the nature of their lower intuitive accuracy remains unclear. Does it reflect a general deficit in intuitive (“System 1”) reasoning, or simply more limited task practice? We addressed this by manipulating task familiarity in two studies with 7th-grade adolescents (approximately 12 years old). We contrasted performance on numerical reasoning tasks involving well-practiced logico-mathematical principles for these students (Study 1: arithmetic word problems; Study 2: decimal comparison) with a task involving a less-practiced principle (the bat-and-ball problem). Using a two-response paradigm, participants first provided an answer under time pressure and cognitive load, followed by a final, unconstrained answer to distinguish intuitive from deliberate reasoning. Across both studies, correct responses on well-practiced tasks were predominantly intuitive, whereas both intuitive and deliberative accuracies were low on the less-practiced task. This supports the view that correct intuitive reasoning depends on prior knowledge automatization and provides evidence against a generic “System 1” developmental deficit.

1. Introduction

Human reasoning is often described as biased. People frequently rely on quick, effortless impressions that can conflict with logical or probabilistic principles [1,2]. According to dual-process theories, such errors arise because two types of processes jointly shape reasoning: intuitive “System 1” processes, which are fast and automatic, and deliberative “System 2” processes, slower and more effortful [3]. In this framework, biases arise when an intuitive but incorrect response prevails and deliberation fails to correct it, so correct reasoning is typically attributed to the successful engagement of deliberate, analytical thinking.

While early dual-process theories portrayed intuition as a source of bias, an accumulating body of evidence shows that our intuitions are not inherently flawed; individuals often generate correct responses from the outset, without the need for further deliberation. In that respect, the two-response paradigm [4], which distinguishes between fast, intuitive answers and slower, deliberative ones, provides direct support for the effectiveness of intuition. In this paradigm, participants are asked to provide two answers to each problem: an initial response under time pressure and cognitive load (to constrain the engagement of deliberation), followed by a second, unconstrained response (to allow for deliberation [5]). Across a range of reasoning tasks, most correct final answers were already correct at the initial, intuitive stage [5–8]. These findings challenge the classical view that deliberation is required to override faulty intuitions and instead suggest that intuitive processes themselves can sometimes produce sound reasoning [4,9–12].

This raises two central questions for dual-process theories: under what conditions do correct logical intuitions emerge, and what mechanisms support their development? One influential proposal is that accurate intuitions emerge through automatization. Stanovich [13] described mindware as the rules, principles, and strategies that individuals acquire through learning and practice. Building on this idea, and consistent with broader evidence that cumulative practice supports academic performance [14,15], De Neys and Pennycook [10] proposed that repeated engagement with logico-mathematical principles allows these principles to become instantiated—i.e., encoded as retrievable mindware—and, with sufficient practice, automatized—i.e., that encoded mindware is retrieved rapidly and with minimal effort. As a result, once-effortful reasoning principles can be retrieved and applied intuitively, without deliberate reflection. When the necessary knowledge has

* Corresponding author.
E-mail addresses: boissinesther@gmail.com (E. Boissin), laura.charbit@u-paris.fr (L. Charbit), elora.taieb@u-paris.fr (E. Taieb), gregoire.borst@u-paris.fr (G. Borst), wim.de-neys@u-paris.fr (W. De Neys).

https://doi.org/10.1016/j.tine.2026.100295
Received 28 January 2026; Received in revised form 15 July 2026; Accepted 23 July 2026
Available online 30 July 2026

been well practiced, it can operate automatically and guide reasoning from the outset, yielding correct responses that are immediate and effortless.

Crucially, multiple intuitions can co-occur; therefore, different candidate responses may be activated in parallel, with the final output reflecting the one that is strongest or most accessible. When the relevant logico-mathematical knowledge is well entrenched, the logical intuition may be activated rapidly and with high accessibility, increasing its chances of dominating over competing incorrect intuitions. In contrast, when logico-mathematical knowledge is less entrenched, incorrect responses should be more likely to prevail. Accordingly, the automatization account entails that correct intuitive responses should be more frequent when the underlying principles have been practiced extensively and can be retrieved automatically.

Preliminary evidence is consistent with this idea. Raoelison et al. [16] and Charbit et al. [17] found that older adolescents (French 12th-grade students, approximately 17 years old) were more likely than early adolescents (French 7th graders, approximately 12 years old) to produce correct intuitive answers on classic reasoning problems such as the bat-and-ball task ("A bat and a ball together cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?"; [18]). Together, these findings suggest that sound intuitive reasoning becomes more prevalent by the end of secondary education in late adolescence, potentially reflecting increased exposure and practice with the underlying logico-mathematical principles [9,10].

At the same time, these results leave open whether age differences truly reflect the effect of automatization or a broader developmental limitation. That is, the precise nature of the less prevalent sound intuiting among young adolescents is not clear. One possibility is that in general, younger reasoners simply have a less efficient "System 1" and are overall less able to generate accurate intuitive responses. Alternatively, the poorer performance originates results from the nature of the tasks typically used in prior dual-process research. Many of the classic reasoning tasks—such as the bat-and-ball problem—draw on principles that are only introduced and practiced during secondary education (e.g., solving an equation with two unknowns, [19]). Younger students are therefore unlikely to have automatized these principles, making correct intuitive responses improbable. From this perspective, the developmental pattern may not indicate a lack of intuitive reasoning capacity but rather limited opportunity to acquire and automatize the relevant knowledge. Distinguishing between these explanations is crucial for understanding how logical intuitions develop. If automatization is the key mechanism, correct intuitions should appear even in younger students—provided the task relies on well-learned, highly practiced principles. Conversely, if a genuine developmental constraint limits the ability to generate correct intuitions, younger students should show poor intuitive performance regardless of how well the underlying principle has been practiced.

The present study tested whether differences in intuitive accuracy among adolescents reflect automatization or broader developmental limitations by focusing on early adolescents (French Grade 7 students, approximately 12 years old) and varying task familiarity—defined as how well the underlying logico-mathematical principle has been practiced—as a proxy for automatization[1]. Two studies used the two-response paradigm to measure both intuitive and deliberative reasoning. In both, we compared performance on a less-practiced task (the bat-and-ball problem, adopted from previous developmental work: [17,20]) with performance on a well-practiced numerical reasoning task that relies on principles extensively taught before the start of secondary education.

In Study 1, we used arithmetic word problems [21] as the well-practiced task. These problems require students to apply basic operations and interpret relational statements in verbal form—skills introduced and reinforced throughout primary school [22,23]. A typical bias occurs when children add rather than subtract, misled by the phrasing "more than." For instance, in "Mary has 8 marbles. She has 5 more marbles than John. How many marbles does John have?"; students often respond 13 instead of 3 ([21], as cited in [22]).

In Study 2, we assessed students' understanding of decimal magnitudes using a decimal comparison task which is typically also introduced and practiced during elementary education [24,25]. A well-documented bias here is the "whole number bias," where students assume that numbers with more digits are larger. For example, they may judge 1.45 to be greater than 1.5 simply because 45 is numerically greater than 5 [26].

If the automatization hypothesis is correct, early adolescents should show higher accuracy on well-practiced tasks than on less practiced tasks at both the intuitive and deliberative stages. Crucially, correct responses should predominantly emerge at the intuitive stage, indicating that well-practiced principles can be applied automatically without deliberation. Conversely, if poorer performance reflects a general developmental limitation, intuitive accuracy should remain low across all tasks, regardless of prior practice.

2. Study 1

2.1. Methods

2.1.1. Preregistration

The design and research questions were pre-registered on the AsPredicted website (https://aspredicted.org) and stored on the Open Science Framework (https://osf.io/6bgdp/overview?view_only=32f712aeef2843b29d05cfaa586a4703) where all data and material can be accessed. No specific statistical analyses were pre-registered.

2.1.2. Participants

In total, we recruited 100 French students in 7th grade (46 women and 10 non-binary, M age = 12.25 years, SD = 0.52 years). A sensitivity analysis indicated that our sample size would allow us to detect small to medium size accuracy differences (effect size dz = .36) between the tasks with 80% power. Data collection occurred in their classrooms after obtaining parental and participants informed consent, during which we explained the research objectives and assured confidentiality in data handling. The study complied with applicable laws and institutional

[1] Throughout the manuscript we use practice and task familiarity interchangeably, although the two are not strictly identical: practice denotes repeated engagement with a principle, whereas familiarity denotes the subjective sense of recognition, and a problem format may feel familiar even when its underlying principle has been little practiced. In the present studies, however, we have no measure of familiarity independent of practice: the tasks were selected on the basis of curricular practice level, such that the well-practiced tasks are also the more familiar ones. The two therefore coincide operationally here. Both remain distinct from automatization, the inferred change in retrieval dynamics that practice is presumed to produce. Although we do not measure automatization directly, practice offers an observable and theoretically grounded index of it: the more a principle has been practiced, the more likely it is to have become automatized.

guidelines and with the Declaration of Helsinki, and was approved by the CER U-Paris (No. 2019-81).

2.1.3. Materials

Participants completed two types of problems in a randomized order: One involving a well-practiced task, the simple arithmetic word problems, and one involving a less-practiced task, the bat-and-ball problem. All problems are provided in Supplementary Material Section A.

Arithmetic word problems. Participants were presented with French arithmetic word problems adapted from Lubin et al. [22]. Each problem was presented sequentially as a set of three sentences: (a) the number of objects of one protagonist (e.g., "Mary has 30 marbles"), (b) the number of objects one protagonist had in excess relative to the other protagonist (e.g., "She has 5 more marbles than John") and (c) the problem's question (e.g., "How many marbles does John have?"). All the operations involved numbers that were all multiples of 5.

Participants had to select a response among four response choices consisting of (1) the correct response (i.e., "25") that consists of the subtraction of both number of objects, (2) the intuitively cued "heuristic" response (i.e., "35") that consists of the addition of both number of objects, (3) a foil option which was the sum of correct and heuristic responses (i.e., "60"), and (4) a second foil option which was the second greatest common divisor multiple of 5 (i.e., "5"). The four response choices appeared in a random order. For example, the arithmetic word problem stated:

(a) *Mary has 30 marbles.*
(b) *She has 5 more marbles than John.*
(c) *How many marbles does John have?*
  ○ 25
  ○ 35
  ○ 60
  ○ 5

Half of the problems were featured in their standard "conflict" version in which the intuitively cued "heuristic" response cues an answer that conflicts with the correct answer. To ensure that participants were engaged in the task, we also presented problems which were featured in their "no-conflict" version, and which were used as control problems. In these control problems, we changed the protagonist for whom the participants must count the marbles. Thus, for the no-conflict items, applying an addition cues the correct answer [27], for instance:

(a) *Mary has 30 marbles.*
(b) *John has 5 more marbles than her.*
(c) *How many marbles does John have?*
  ○ 35
  ○ 20
  ○ 55
  ○ 10

In this case the intuitively cued "35" answer is correct. The four response options for the no-conflict items were: (1) the correct response ("35"), based on the addition of the two quantities; (2) a first foil ("20"), corresponding to half the correct response, rounded to the nearest multiple of 5; (3) a second foil ("55"), obtained by summing the correct response and the first foil; and (4) a third foil ("10"), corresponding to the largest multiple of 5 shared by foils 1 and 2, excluding their values.

Overall, these control no-conflict 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.

Two sets of problems were used in order to counterbalance problem content: The conflict problems in one set were the no-conflict problems in the other, and vice-versa. The presentation order of the conflict and no-conflict problems was randomized in each set. Participants were randomly assigned to one of the two sets for each block and were presented in total with five conflict and five no-conflict items.

Bat-and-ball problems. Participants were also presented with bat-and-ball problems taken from Raoelison and De Neys [8], translated from English to French. They were modified versions of the original 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 [6–8]. Bat-and-ball problems were presented serially, with the sentences introduced from (a) to (c) in that order. Participants had to select the correct response among four response choices consisting of (1) the correct response (i.e., which would be "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 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 [2]. 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:

(a) *In a company, there are 150 men and women in total.*
(b) *There are 100 more men than women.*
(c) *How many women are there?*
  ○ 25
  ○ 50
  ○ 75
  ○ 10

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 [28,29], for instance:

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

In this case the intuitively cued "50" answer is also correct. 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 [8]. Two sets of problems were used in order to counterbalance problem content: The presentation order of the conflict and no-conflict problems was randomized in each set. Participants were randomly assigned to one of the two sets for each block and were presented in total with five conflict and five no-conflict items.

Two-response format. For both the arithmetic word and the bat-and-ball problems, participants responded to each problem using a two-response procedure, where they first provided a "fast" answer, directly followed by a second "slow" answer [4]. 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 [5,6,8]). The load task was based on the dot memorization task [30] given that it had been successfully used to burden executive resources during reasoning tasks [31,32]. Participants had to memorize a complex visual pattern (i.e., 4 crosses in a 3 × 3 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 [6], for more details).

As in Bago and De Neys [6], a time limit of 5 seconds was chosen for the initial response of the bat-and-ball problems, based on pretesting that indicated it simply amounted to the time needed to read the preambles, move the mouse, and select an answer among the four possibilities (see [6] for details).

Note that these time and load constraints are intended to limit the contribution of slow, deliberate processing to the initial response. It remains possible, however, that some deliberation is rapid rather than slow; whether such fast processing should be classified as intuitive or as fast deliberation is a distinct, orthogonal question of definition [9,37]. We therefore treat the initial response as intuitive in an operational sense—defined by the time and load constraints—independently of this broader definitional debate.

Since the amount of information presented on screen is highly similar in the arithmetic word problems and bat-and-ball problems (i.e., same number of sentences and approximately the same number of words), we decided to use the same exact deadline for the initial response stage of the arithmetic word problems. Participants were thus asked to give their first intuitive response within a maximum of 5

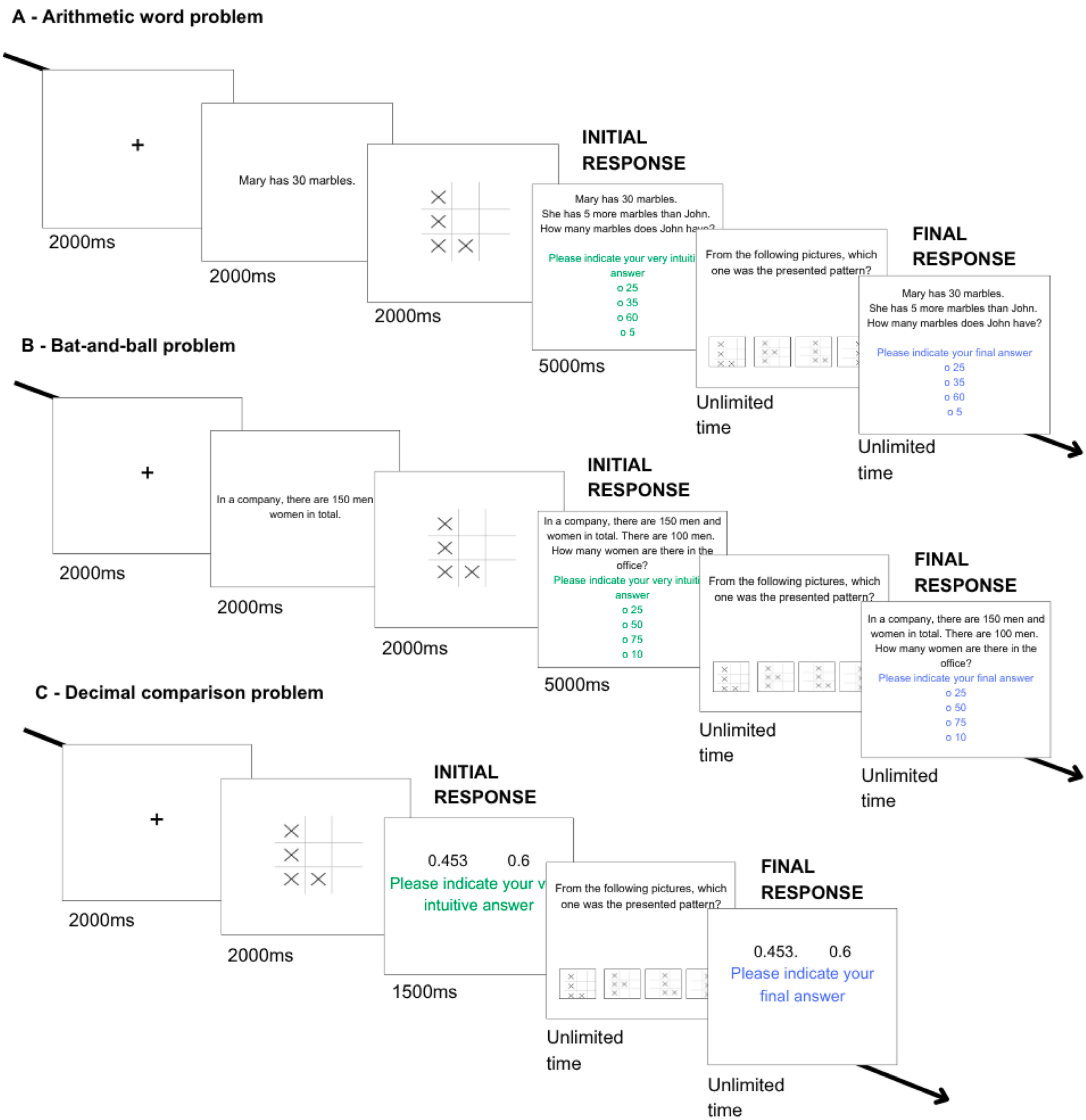

Fig. 1. Time course of a complete two-response trial for (A) arithmetic word problem in Study 1, (B) the bat-and-ball problem in Study 1 and 2, and (C) the decimal comparison problem in Study 2.

seconds. It is worth noting that Lubin et al. [27] showed that 6th graders (approximately 11 years old) solve the arithmetic word problems in an average of 10 seconds, indicating that our 5-second deadline is highly challenging for 7th graders (approximately 12 years old).

In addition, participants are also under secondary task load when giving their initial response both for the arithmetic word and bat-and-ball tasks. 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 demanding, particularly for younger participants. Early adolescents may struggle to read and process problems under time pressure and memory load. However, prior studies [16,17,20] have shown that this procedure can be reliably used with students as young as 12 years old, using the same time constraints and memorization load to both younger (≈12 years old) and older adolescents (≈17 years old).

2.1.4. Procedure

The study was run individually on iPads. The participants were tested in their classrooms, in groups, under the supervision of at least one teacher. Participants were instructed that the study would take thirty minutes and that it demanded their full attention. The same procedure was applied regardless of whether participants started by one or the other task. For both problems, a general description of the task was presented in which participants were instructed that they would need to solve reasoning problems, for which they would have to provide two consecutive responses. They were told that we were interested in their very first, initial answer that comes to mind and that –after providing their initial response –they could reflect on the problem and take as much time as they needed to provide a final answer. In order to familiarize themselves with the two-response procedure, they first solved two unrelated practice problems. Next, they familiarized themselves with the cognitive load procedure by solving two load trials and, finally, they solved two problems which included both cognitive load and the two-response procedure.

Fig. 1 shows a typical trial both for arithmetic word and bat-and-ball problems, which started with a fixation cross for 2000ms, followed by the first sentence of the problem (e.g., "In an office, there are 150 men and women in total" or "Mary has 30 marbles.") for 2000ms, and followed by the visual matrix for the cognitive-load task for 2000ms. Then the full problem was presented, at which point participants had 5000ms to select their initial answer among the four choices. After 3000ms the background of the screen turned yellow to warn participants that they only had a short amount of time left to answer. If they 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, participants were presented with four visual matrices and had to choose the one that they had previously memorized. They received feedback as to whether their 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 problem was presented again, and participants were asked to provide a final deliberate answer (with no time limit). Task order was randomized across participants. At the end of the study, participants were asked to complete a page with demographic questions.

Response times were recorded at both response stages. A detailed analysis of response times falls outside the scope of the present paper; mean final response times on conflict items, broken down by task and accuracy, are reported in the Supplementary Material Section B for the interested reader.

2.1.5. Trial exclusion

We discarded trials in which participants failed to provide their initial response before the deadline (7.9% of all trials of bat-and-ball problems and 17.9% of all trials of arithmetic word problems) or failed to pick the correct matrix in the load task (23% of the remaining trials of bat-and-ball problems and 27.8% of the remaining trials of the arithmetic word problems), and we analyzed the remaining 70.9% of all bat-and-ball trials and 59.3% of all arithmetic word trials. On average, for the arithmetic word problems, each participant contributed 3.27 ($SEM = 0.14$) conflict trials, and 2.66 ($SEM = 0.12$) no-conflict trials out of 5. For the bat-and-ball trials, on average, each participant contributed to 3.37 ($SEM = 0.13$) conflict trials and 3.72 ($SEM = 0.12$) no-conflict trials out of 5. To ensure that trial exclusion did not drive our results, Supplementary Material Section C reports exclusion rates broken down by task and item type (i.e., conflict vs. no-conflict items), tests of whether exclusion differed across tasks, and two robustness checks—one retaining all excluded trials and coding them as incorrect, the other re-analyzing the data under a more lenient inclusion criterion that excluded only trials with a missed deadline. In every case the main pattern of results was unchanged across both studies, indicating that the task differences are not attributable to differential exclusion.

2.2. Results

Response accuracy. We first compared accuracy on the well-practiced arithmetic word task and the less-practiced bat-and-ball task, separately for initial and final responses on conflict trials.

For each response stage, we fitted a logistic mixed-effects model predicting accuracy from task type (arithmetic vs. bat-and-ball), with random intercepts for participants to account for repeated measures.

Fig. 2 shows the results. Final conflict accuracy was substantially lower for the less-practiced bat-and-ball task ($M = 4.82\%$, $SEM = 1.96\%$) than for the well-practiced arithmetic word task ($M = 47.49\%$, $SEM = 4.37\%$), $OR = 143.96$, 95% CI [54.28, 381.79], $p < .001$.

The same pattern emerged for initial intuitive responses. Accuracy was significantly higher for the well-practiced arithmetic word task ($M = 46.70\%$, $SEM = 3.94\%$) than for the less-practiced bat-and-ball task ($M = 6.16\%$, $SEM = 1.95\%$), $OR = 31.12$, 95% CI [15.64, 61.91], $p < .001$. Together with the final response results, this supports the view that both intuitive and deliberative accuracy are affected by task practice.

As predicted, for the no-conflict control problems, we observed that performance was very high, both for the well-practiced arithmetic word task (M *initial response* $= 69.68\%$, $SEM = 3.67\%$, M *final response* $= 88.80\%$, $SEM = 2.63\%$) and for the less-practiced bat-and-ball task (M *initial response* $= 91.16\%$, $SEM = 2.53\%$, M *final response* $= 94.50\%$, $SEM = 2.00\%$). This high performance indicates that participants managed to process the problems and were not randomly guessing. If the initial response constraints prevented participants from reading the problems and processing the basic problem information, we should observe more random responses.

Direction of change. To better understand how participants changed (or did not change) their responses after deliberation, we performed a direction of change analysis for the conflict items [5]. Especially, each trial consisted 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. Fig. 3 shows the direction of change distribution for each task.

First, as Fig. 3 shows, early adolescents predominantly gave biased "00" responses on the less-practiced bat-and-ball task ($M = 91.18\%$, $SEM = 2.39\%$), but less so on the well-practiced arithmetic task ($M = 50.86\%$, $SEM = 4.34\%$). Critically, when participants answered correctly on the well-practiced arithmetic word task, they typically did so at the intuitive stage, as reflected by a higher proportion of “11” response patterns ($M = 31.77\%$, $SEM = 3.88\%$) compared to "01" response patterns ($M = 9.18\%$, $SEM = 1.81\%$). A logistic mixed-effects model restricted to the well-practiced task, contrasting “11” with “01” responses and including a random intercept by participant, showed a significant positive effect, OR

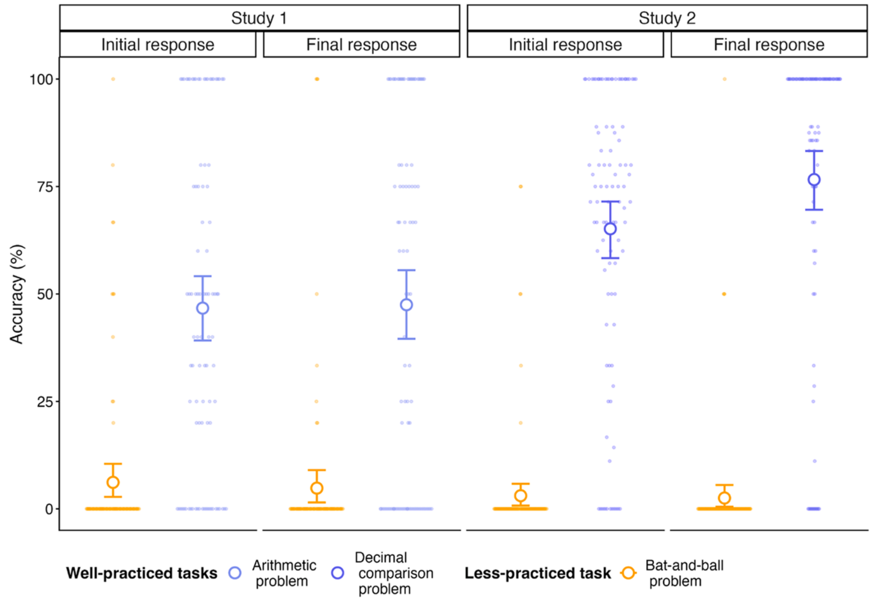

Fig. 2. Mean initial and final accuracy on conflict problems in Study 1 (less-practiced bat-and-ball vs. well-practiced arithmetic word problems) and Study 2 (less-practiced bat-and-ball vs. well-practiced decimal comparison). Error bars represent standard errors of the mean (SEM).

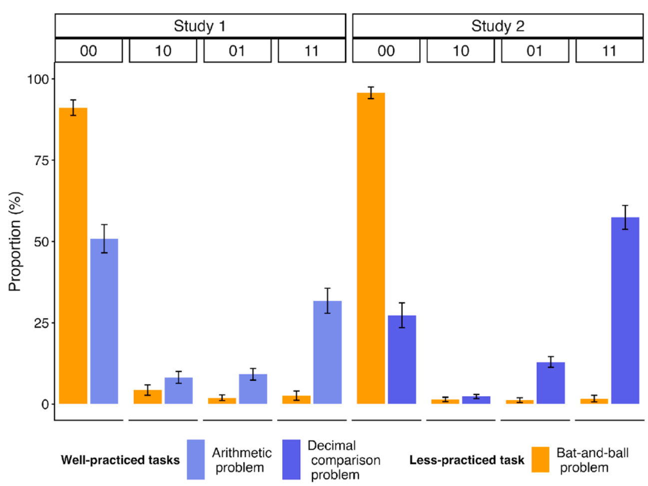

Fig. 3. Proportion of each direction of change (i.e., "00," "10," "01" and "11" response patterns) for the conflict problems for Study 1 (less-practiced Bat-and-ball vs. well-practiced Arithmetic word tasks) and for Study 2 (the less-practiced Bat-and-ball vs. well-practiced Decimal comparison tasks). Error bars represent standard errors of the mean (SEM).

= 3.56, 95% CI [2.06, 6.15], $p < .001$. This suggests that correct responses on the well-practiced task often emerged intuitively and did not depend on later correction.

For the less-practiced bat-and-ball task, correct responses were rare, and there was no significant difference between those given at the initial intuitive stage (*M "11" response patterns* = 2.58%, SEM = 1.45%) and those given only after deliberation (*M "01" response patterns* = 1.93%, SEM = 0.87), *OR* = 1.32, 95% CI [0.30, 5.75], $p = .714$. Overall, these findings suggest that correct responses on the well-practiced arithmetic word task were primarily driven by accurate intuition, whereas early adolescents struggled to solve the less-practiced bat-and-ball problem—even after deliberation.

3. Study 2

In Study 2, we aimed to generalize the effects of practice to a different type of task involving a different well-practiced logico-mathematical principle: decimal magnitude comparison. Hence, early adolescents completed two types of tasks in randomized order: a well-practiced decimal comparison task and a less-practiced task, once again the bat-and-ball. In the decimal comparison task, participants had to judge which of two decimal numbers is greater. The bat-and-ball problems were similar to those used in Study 1. All problems are detailed in Supplementary Material Section A.

3.1. Methods

3.1.1. Preregistration

The design and research questions were pre-registered on the AsPredicted website (https://aspredicted.org) and stored on the Open Science Framework (https://osf.io/6bgdp/overview?view_only=32f712aeef2843b29d05cfaa586a4703) where all data and material can be accessed. No specific statistical analyses were pre-registered.

3.1.2. Participants

In total, we recruited 100 French students in 7th grade (56 women and 8 non-binary, *M age* = 12.34 years, *SD* = 0.55 years). Data collection occurred in their classrooms after obtaining parental and participants informed consent, during which we explained the research objectives and assured confidentiality in data handling. The study complied with applicable laws and institutional guidelines and with the Declaration of Helsinki, and was approved by the CER U-Paris (No. 2019-81).

3.1.3. Materials

Decimal comparison problems. Participants were presented with decimal comparison problems adapted from Roell et al. [33]. On each trial, a pair of decimal numbers was shown, and participants were asked to select the larger one. The smaller number always had more digits after the decimal point—for example, "0.453 vs. 0.6." In such cases, the common heuristic error is to choose "0.453," based on the incorrect belief that longer decimals represent larger values. However, the correct answer is "0.6."

The decimal numbers included one, two, or three digits after the decimal point. Task difficulty was systematically varied by manipulating the numerical distance between the first digits following the decimal point, with distances ranging from 1 to 6.

Two sets of problems were used to counterbalance item content: conflict problems in one set served as no-conflict problems in the other, and vice versa. The order of conflict and no-conflict trials was randomized within each set. Participants were randomly assigned to one of the two sets per block and completed ten conflict and ten no-conflict items in total.

Two-response format (and development). As in Study 1, participants completed the decimal comparison task using a two-response procedure: an initial fast, intuitive response followed by a final slow, deliberate response [4]. Given the different task format of the bat-and-ball and decimal task, the response deadline for the intuitive stage was set at 1.5 seconds, based on a pilot study with 33 French 7th-grade students (*M age* = 12.19 years, *SD* = 0.47; 16 female, 4 non-binary). In this pilot, students gave correct responses with an average response time of 1.53 seconds (*SEM* = 0.10) on conflict items. We decided to set the absolute deadline threshold to 1.5 seconds (and not lower) to create time-pressure while keeping the number of missed trials acceptable (see Trial Exclusions).

3.1.4. Procedure

The study was run individually on iPads. The participants were tested in their classrooms, in groups, under the supervision of at least one teacher. Participants were instructed that the study would take thirty minutes and that it demanded their full attention. For the decimal comparison tasks, participants were instructed to determine which of the two decimal numbers was the largest. First, a general description of the task was presented in which participants were instructed that they would need to solve decimal comparison problems, for which they would have to provide two consecutive responses. They were told that we were interested in their very first, initial answer that came to mind and that –after providing their initial response –they could reflect on the item and take as much time as they needed to provide a final answer. In order to familiarize themselves with the two-response procedure, they first solved two unrelated practice problems. Next, they familiarized themselves with the cognitive load procedure by solving two load trials and, finally, they solved two items which included both cognitive load and the two-response procedure.

Fig. 1 shows a typical trial for the decimal comparison task which started with a fixation cross for 2000ms, followed by the visual matrix for the cognitive-load task for 2000ms and followed by the pairs of decimals (e.g., "0.453 vs. 0.6") where participants had up to 1500 ms to select their initial response. If they 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, participants were presented with four visual matrices and had to choose the one that they had previously memorized. They received feedback as to whether their 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 bat-and-ball task was presented as in Study 1. Task order was randomized across participants. At the end of the study, participants were asked to complete a page with demographic questions.

Response times were recorded at both response stages. A detailed analysis of response times falls outside the scope of the present paper; mean final response times on conflict items, broken down by task and accuracy, are reported in the Supplementary Material Section B for the interested reader.

3.1.5. Trial exclusion

We discarded trials in which participants failed to provide their initial response before the deadline (20.3% of all trials of decimal comparison problems and 6.9% of all trials of bat-and-ball problems) or failed to pick the correct matrix in the load task (17.6% of the remaining trials of decimal comparison problems and 21.6% of the remaining trials of the bat-and-ball problems), and we analyzed the remaining 65.2% of all decimal comparison trials and 73.0% of all bat-and-ball trials. On average, for the decimal comparison problems, each participant contributed 6.29 (*SEM* = 0.21) conflict trials, and 6.76 (*SEM* = 0.22) no-conflict trials out of 10. For the bat-and-ball trials, on average, each participant contributed to 3.48 (*SEM* = 0.11) conflict trials and 3.82 (*SEM* = 0.11) no-conflict trials out of 5. As in Study 1, trial exclusion did not drive our results (see Supplementary Material Section C).

3.2. Results

Response accuracy. Fig. 2 presents the results. As in Study 1, final response accuracy was significantly higher for the well-practiced decimal comparison task (*M* = 76.61%, *SEM* = 3.62%) than for the less-practiced bat-and-ball task, where performance was notably low (*M* = 2.53%, *SEM* = 1.32%), *OR* = 11,764.07[2], 95% CI [2307.03, 59987.75], $p < .001$. The same pattern emerged for initial intuitive responses, with higher accuracy on the well-practiced decimal task (*M* = 66.94%, *SEM* = 3.49%) compared to the less-practiced bat-and-ball task (*M* = 3.06%, *SEM* = 1.32%), *OR* = 297.87, 95% CI [127.52, 695.79], $p < .001$. As in Study 1, participants performed substantially better on the well-practiced task, even at the intuitive stage.

No-conflict accuracies for both tasks and each response stage were very high for final responses (M well-practiced decimal comparison task = 92.61%, *SEM* = 2.53%; M less-practiced bat-and-ball task = 95.86%, *SEM* = 1.82%) and initial responses (M well-practiced decimal comparison task = 88.93%, *SEM* = 2.56%; M less-practiced bat-and-ball task = 92.95%, *SEM* = 1.87%).

Direction of change. For the well-practiced decimal comparison task, participants most frequently provided correct responses at both the initial and final stages ("11" response patterns; *M* = 59.02%, *SEM* = 3.78%), while incorrect responses at both stages ("00" response patterns) were less common (*M* = 27.30%, *SEM* = 3.81%). Additionally, participants gave significantly more "11" responses compared to responses that were only correct at the deliberative stage ("01" response patterns; *M* = 11.36%, *SEM* = 1.66%), *OR* = 6.86, 95% CI [4.44, 10.61], $p < .001$. This suggests that when participants respond correctly in the well-practiced decimal comparison task, they are able to do so intuitively, without requiring further deliberation.

In contrast, for the less-practiced bat-and-ball task, participants predominantly gave incorrect responses at both stages ("00" response patterns; *M* = 95.73%, *SEM* = 1.81%). Correct responses were rare, whether at the intuitive stage ("11" patterns; *M* = 1.75%, *SEM* = 1.09%) or only after deliberation ("01" patterns; *M* = 1.13%, *SEM* = 0.68%), with no significant difference between the two, *OR* = 1.25, 95% CI [0.34, 4.65], $p = .739$. These findings confirm that early adolescents were more successful on the well-practiced task and that correct responses primarily emerged at the intuitive stage.

4. Discussion

These two studies tested whether automatization could account for the emergence of (in)correct intuitive reasoning in early adolescence. By focusing on early adolescents and varying task familiarity, we found that they were highly accurate on well-practiced numerical tasks, with correct responses predominantly given at the intuitive stage, whereas performance on less-practiced tasks remained poor even after deliberation. These findings are consistent with the automatization idea: when logico-

[2] Because accuracy on the less-practiced bat-and-ball task was near floor, the resulting odds ratios are extremely large and reflect a situation of near-complete separation in the model rather than a meaningful quantitative effect size. This does not weaken our conclusion, however, as it precisely reflects the massive performance gap between the well-practiced and the less-practiced tasks predicted by our hypothesis.

mathematical mindware has been instantiated and sufficiently automatized through repeated practice, it can be retrieved rapidly and guide sound reasoning from the outset. This pattern suggests that younger adolescents are not limited by a general deficit in intuitive “System 1” reasoning but by the extent to which relevant principles have been practiced and automatized. The emergence of sound intuitions therefore appears to arise from prior exposure and knowledge automatization, not from the gradual lifting of a general developmental constraint on intuitive processing.

Inherently, our results also question the alleged role of deliberation in sound reasoning. They suggest that deliberation may not always be necessary—or even effective—in correcting intuitive errors, particularly when the underlying principles are already well practiced and automatized. These findings do not show deliberation to be generally ineffective, however. On the less-practiced bat-and-ball task, unconstrained final responses did not substantially improve accuracy, which remained near floor at both stages. On the well-practiced tasks, deliberation did produce some additional correct responses—most notably on the decimal comparison task, where final accuracy exceeded descriptively initial accuracy—yet correct responses still emerged predominantly, though not exclusively, at the intuitive stage. In the present studies and with this population, deliberation was thus neither the main source of correct responding when the relevant principles were well practiced nor an effective remedy when they were not. Although our data do not directly address this, the pattern is consistent with emerging arguments [34–36] that deliberation may serve more to justify responses than to improve accuracy, functioning as a post-hoc rationalization rather than a genuine correction mechanism. From this perspective, and at least for the tasks studied here, repeated practice that supports automatization may be a more effective route to sound reasoning than the active engagement of deliberation. The precise role and functions of the deliberative system, however, remain unclear [37].

The role and functions of the deliberative system are rather important given that the cognitive control on which deliberation is presumed to depend operates under sharply limited resources. Musslick and Cohen [38] argue that this scarcity forces a trade-off: the controlled processes that allow us to commit fully to a single task cannot, at the same time, handle the multiple competing demands that we are consistently facing in our daily lives. Cognitive control thereby acts as a bottleneck on the system, and this bottleneck makes lighter, less costly processes not merely useful but necessary. Automatization, in this view, is not only an outcome of practice but a need of bypassing the constraints imposed by the limited cognitive control on which deliberation relies.

Yet if automatization is so central to reasoning, when and how it actually develops remains underexplored. Neither the amount of practice nor the quality of the learning conditions required to bring it about is well established. Large-scale modeling of academic learning suggests that students typically need about seven practice opportunities per knowledge component to reach mastery (defined as 80% correctness), provided that practice is accompanied by immediate feedback on errors and contextual instruction [14]. Whether a comparable threshold characterizes the development of correct intuitive responding on conflict reasoning tasks is itself an open question, but the route to automatization may depend at least as much on the qualitative structure of practice as on its sheer quantity. Identifying the conditions that favor automatization—and, in turn, the generation of correct intuitions in reasoning—is therefore an important direction for future research.

Nonetheless, several limitations must be acknowledged. First, our inference about automatization and practice is indirect. We do not directly measure automatization, but infer it from presumed curricular exposure. Although this inference is grounded in prior literature, future studies may want to include more direct testing of automatized processing (e.g., by observing the effects of extended in-lab training, [39]).

A related limitation concerns the comparability of the tasks themselves. Practice level is not the only dimension on which the tasks vary, and we cannot fully rule out that some of the observed performance differences could reflect surface task properties or intrinsic differences in difficulty, rather than—or in addition to—differences in the degree of automatization. Indeed, the well-practiced and less-practiced tasks were not closely matched on all relevant dimensions: in Study 1, arithmetic word problems and bat-and-ball problems differ in relational wording and problem structure, and in Study 2, decimal comparison and bat-and-ball problems differ in representational format—one involving a numerical comparison, the other a verbal problem with a misleading relational statement. One might also worry that the bat-and-ball task is simply more difficult than the well-practiced tasks, independently of practice. From an automatization perspective, however, difficulty and practice are not fully separable: a task is cognitively demanding precisely to the extent that the underlying principles have not been sufficiently practiced and automatized. This relation is central to the mindware framework on which we rely [13]. The greater difficulty of the bat-and-ball task may therefore be a consequence of its lower practice level rather than an independent dimension of variation. In any case, the no-conflict data provide some reassurance: performance was uniformly high on no-conflict versions of all tasks, including the bat-and-ball, suggesting that neither the structural properties nor the intrinsic difficulty of the less-practiced task in themselves prevent correct responding. The performance gap therefore appears to be specific to conflict items, which is more consistent with a practice-based account than with a generic task-format or task-difficulty account. Future research should nonetheless aim to better match tasks on representational demands, verbal complexity, and problem structure, so that practice level can be more cleanly isolated as the critical variable.

A further potential concern, specific to Study 2, bears on the source of systematic error in the decimal comparison task. In Study 2, we treated the whole-number bias—the tendency to judge decimals with more digits as larger—as the primary heuristic driving incorrect responses, consistent with the established literature [24–26]. Following a reviewer's suggestion, we also examined two alternative strategies that students could in principle use: comparing the rightmost digit across the two decimals, or comparing the largest individual digit appearing in each decimal. We reclassified each item pair as conflict or no-conflict separately for each candidate strategy and recomputed accuracy at both response stages (full details are reported in the Supplementary Material Section D). The rationale is straightforward: if students rely on a given heuristic, accuracy should be higher on no-conflict items—where the heuristic points to the correct response—than on conflict items, where it is misleading. The size of the gap between no-conflict and conflict accuracy can therefore be taken as an index of how strongly a given heuristic shapes responses, with a larger gap indicating that students are more likely to rely on that heuristic when they err. Under the rightmost-digit classification, accuracy did not differ between conflict and no-conflict items, indicating no detectable reliance on this strategy. The largest-digit classification did produce a conflict effect—accuracy was lower on conflict than on no-conflict items—but the gap was markedly smaller than the one obtained under the whole-number bias. The two classifications are not fully independent—items counted as conflict under the largest-digit strategy form a subset of whole-number-bias conflict items—so we cannot determine whether this residual effect reflects genuine reliance on a largest-digit strategy or merely the underlying whole-number bias. This ambiguity, however, does not affect our interpretation. The whole-number bias yields a more pronounced conflict effect than any alternative classification, indicating that it is the heuristic that best captures students' systematic errors—consistent with its prominence in the literature. Whether students rely on the whole-number bias or, more weakly, on a largest-digit strategy, our conclusions are not inflated by the choice of classification: the heuristic we treat as primary is also the one whose impact on responses is most clearly visible in the data. Final-response accuracy on the well-practiced decimal task remained substantially higher than initial accuracy and correct responses emerged predominantly intuitively under either reading.

A broader limitation applies to both studies: we did not measure participants' prior knowledge and abilities directly. We did not collect measures of prior mathematical achievement, classroom exposure, or pretest knowledge of the relevant principles, and our characterization of arithmetic word problems and decimal comparison as well-practiced, and of the bat-and-ball principle as less-practiced, rests on curricular considerations [22–25] rather than on measured exposure. Although this curricular argument is grounded in prior work, direct measures of achievement and exposure would provide stronger evidence and represent a valuable addition for future research. In their absence, our data cannot determine the extent to which the individual differences in intuitive accuracy we observe reflect automatization specifically, as opposed to general mathematical ability, working memory, reading comprehension, or prior instruction. Importantly, these factors are unlikely to be independent of practice: they may shape intuitive accuracy in part through their effect on how readily the relevant principles become instantiated and automatized. Our design therefore does not allow us to isolate their respective contributions, and we treat them as candidate mechanisms operating alongside—or through—automatization rather than as clean alternatives to it.

These individual differences are all the more apparent given that, even on well-practiced tasks, accuracy was far from ceiling. This suggests that automatization is not an all-or-nothing process but a gradual one. Among adolescents, some students may have reached a level of automatization that enables accurate intuitive responding, while others still rely on effortful reasoning or default to misleading heuristics. In other words, some students may have achieved full automatization, while others have not. These gaps likely reflect variability in how quickly principles can be instantiated and automatized. While our data do not directly address these mechanisms, prior research points to several plausible candidates—most notably, cognitive abilities and dispositional factors. One possibility is that people with higher cognitive ability are better able to detect regularities, abstract principles, and consolidate them into stable representations. On this view, cognitive ability may contribute not only to deliberate reasoning but also to the automatization of relevant principles. Supporting this, Charbit et al. [17] found that cognitive ability predicted deliberative performance on the bat-and-ball task among older adolescents (i.e., 12th graders, approximately 17 years old), but not in early adolescents (i.e., 7th graders, approximately 12 years old). Similarly, in adults, cognitive ability mainly predicts intuitive accuracy on the same task [40]. These findings suggest that the link between cognitive ability and reasoning—both intuitive and deliberative—emerges progressively as principles become automatized. Future research should test this directly by manipulating both task automatization and individual cognitive ability.

Dispositional factors may also influence how effectively logico-mathematical principles become automatized. Traits such as actively open-minded thinking have been linked to greater intuitive accuracy and greater responsiveness to debiasing interventions [41]. Although the mechanisms remain to be clarified, such dispositions may promote greater attention to normatively relevant features during learning, increasing the likelihood that correct principles are instantiated and applied automatically. Taken together, these considerations suggest that intuitive accuracy reflects not only the amount of exposure to a principle but also the cognitive and dispositional abilities that shape how effectively that exposure leads to automatization. This possibility remains largely untested and should be a focus for future research.

At the same time, individual variability in intuitive accuracy may not only result from cognitive abilities or dispositions, but also from the quantity and quality of prior practice opportunities. Recent evidence suggests that students learn at remarkably similar rates per practice opportunity but differ substantially in their starting points—that is, in how much of a given knowledge they have already mastered when a new learning episode begins [14]. On this view, the variability observed even on well-practiced tasks may reflect less a difference in how efficiently students can automatize a given principle than a difference in how many opportunities to practice that principle they have already encountered—opportunities that depend on educational backgrounds and access to structured learning experiences. Disentangling these contributions from those of cognitive ability and disposition is a task for future research.

Another limitation concerns the focus on early adolescents of around 12 years old and a narrow range of tasks which limit the broader applicability of these findings. Future research should expand to include a wider range of age groups and developmental stages to explore how automatization evolves over time. Additionally, while the tasks in this study targeted key logico-mathematical principles, they do not cover the full spectrum of reasoning challenges students encounter. Extending this research to areas such as algebra, probabilistic reasoning, or real-world problem-solving could offer a more comprehensive understanding of how automatization influences reasoning across different domains.

Finally, while our findings provide support for the automatization hypothesis, they do not resolve ongoing debates about the underlying nature of the intuitions themselves. Some theorists have argued that correct intuitive responses may not reflect genuine sensitivity to formal logico-mathematical principles or structure, but rather the activation of heuristic cues that simply happen to align with the correct answer (e.g., [42–46]). Our study does not allow us to adjudicate between these two views. However, it does suggest that the reliability of intuition—whatever its underlying source—is shaped by prior exposure and familiarity with relevant problem features. Whether the correct intuitive response reflects genuine logico-mathematical principles or a domain-specific heuristic proxy remains an open empirical question. Yet in both cases, the conclusion is the same: intuition reflects the knowledge structures and response patterns that prior experience has made readily accessible.

In sum, our findings suggest that developing reasoners are not constrained by a general "System 1" deficit but by the extent to which relevant principles have been learned and automatized. Sound intuitions emerge when knowledge has been sufficiently practiced to guide reasoning effortlessly, supporting an automatization account of dual-process development and highlighting knowledge automatization as a key driver of reasoning accuracy.

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.

Ethical statement

The study complied with applicable laws and institutional guidelines and with the Declaration of Helsinki and was approved by the CER U-Paris (No. 2019-81). Written informed consent was obtained from all participants and from their parents/guardians.

CRediT authorship contribution statement

Esther Boissin: Writing – review & editing, Writing – original draft, Visualization, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Laura Charbit:** Writing – review & editing, Writing – original draft, Investigation. **Elora Taieb:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Grégoire Borst:** Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization. **Wim De Neys:** Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors report there are no competing interests to declare.

Acknowledgements

We would like to thank Ana Hoban and Lucas Etheve for their invaluable help with data collection.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tine.2026.100295.

References

[1] J.St.B.T. Evans, Dual-processing accounts of reasoning, judgment, and social cognition, Annu. Rev. Psychol. 59 (2008) 255–278, https://doi.org/10.1146/annurev.psych.59.103006.093629.
[2] D. Kahneman, Thinking, Fast and Slow, Farrar, Straus and Giroux, 2011.
[3] J.St.B.T. Evans, K.E. Stanovich, Dual-process theories of higher cognition: advancing the debate, Perspect. Psychol. Sci. 8 (2013) 223–241, https://doi.org/10.1177/1745691612460685.
[4] V.A. Thompson, J.A. Prowse Turner, G. Pennycook, Intuition, reason, and metacognition, Cognit. Psychol. 63 (2011) 107–140, https://doi.org/10.1016/j.cogpsych.2011.06.001.
[5] B. Bago, W. De Neys, Fast logic?: Examining the time course assumption of dual process theory, Cognition 158 (2017) 90–109, https://doi.org/10.1016/j.cognition.2016.10.014.
[6] B. Bago, W. De Neys, The smart system 1: evidence for the intuitive nature of correct responding on the bat-and-ball problem, Think. Reason. 25 (2019) 257–299, https://doi.org/10.1080/13546783.2018.1507949.
[7] E.M. Janssen, M. Raoelison, W. de Neys, "You're wrong!": the impact of accuracy feedback on the bat-and-ball problem, Acta Psychol. 206 (2020) 103042, https://doi.org/10.1016/j.actpsy.2020.103042.
[8] M. Raoelison, W. De Neys, Do we de-bias ourselves?: the impact of repeated presentation on the bat-and-ball problem, Judgm. Decis. Mak. 14 (2019) 170–178.
[9] W. De Neys, Advancing theorizing about fast-and-slow thinking, Behav. Brain Sci. 46 (2023) e111, https://doi.org/10.1017/S0140525X2200142X.
[10] W. De Neys, G. Pennycook, Logic, fast and slow: advances in dual-process theorizing, Curr. Dir. Psychol. Sci. 28 (2019) 503–509, https://doi.org/10.1177/0963721419855658.
[11] G. Pennycook, J.A. Fugelsang, D.J. Koehler, What makes us think? A three-stage dual-process model of analytic engagement, Cognit. Psychol. 80 (2015) 34–72, https://doi.org/10.1016/j.cogpsych.2015.05.001.
[12] V.F. Reyna, A new intuitionism: meaning, memory, and development in Fuzzy-Trace Theory, Judgm. Decis. Mak. 7 (2012) 332–359, https://doi.org/10.1017/S1930297500002291.
[13] K.E. Stanovich, Miserliness in human cognition: the interaction of detection, override and mindware, Think. Reason. 24 (2018) 423–444, https://doi.org/10.1080/13546783.2018.1459314.
[14] K.R. Koedinger, P.F. Carvalho, R. Liu, E.A. McLaughlin, An astonishing regularity in student learning rate, Proc. Natl. Acad. Sci. 120 (2023) e2221311120, https://doi.org/10.1073/pnas.2221311120.
[15] M.W.H. Spitzer, Just do it! Study time increases mathematical achievement scores for grade 4–10 students in a large longitudinal cross-country study, Eur. J. Psychol. Educ. 37 (2022) 39–53, https://doi.org/10.1007/s10212-021-00546-0.
[16] M. Raoelison, E. Boissin, G. Borst, W. De Neys, From slow to fast logic: the development of logical intuitions, Think. Reason. 27 (2021) 599–622, https://doi.org/10.1080/13546783.2021.1885488.
[17] L. Charbit, E. Boissin, M. Raoelison, W. De Neys, Emergence of the smart intuitor: how cognitive ability shapes adolescent reasoning, Think. Reason. 32 (2025) 70–91, https://doi.org/10.1080/13546783.2025.2527672.
[18] S. Frederick, Cognitive reflection and decision making, J. Econ. Perspect. 19 (2005) 25–42, https://doi.org/10.1257/089533005775196732.
[19] J. Hoover, A. Healy, The bat-and-ball problem: a word-problem debiasing approach, Think. Reason. 27 (2021) 567–598, https://doi.org/10.1080/13546783.2021.1878473.
[20] E. Boissin, L. Charbit, M. Raoelison, G. Borst, S. Caparos, W. De Neys, Reactivating logic? Age-dependent effects of intuitive debiasing training, Think. Reason. (2025) under review.
[21] M.S. Riley, J.G. Greeno, J.I. Heller, Development of children's problem solving ability in arithmetic, in: H.P. Ginsburg (Ed.), The Development of Mathematical Thinking, Academic Press, New York, 1983, pp. 153–196.
[22] A. Lubin, O. Houdé, W. de Neys, Evidence for children's error sensitivity during arithmetic word problem solving, Learn. Instr. 40 (2015) 1–8, https://doi.org/10.1016/j.learninstruc.2015.07.005.
[23] L. Verschaffel, Using retelling data to study elementary school children's representations and solutions of compare problems, J. Res. Math. Educ. 25 (1994) 141–165, https://doi.org/10.2307/749506.
[24] L. Desmet, J. Grégoire, C. Mussolin, Developmental changes in the comparison of decimal fractions, Learn. Instr. 20 (2010) 521–532, https://doi.org/10.1016/j.learninstruc.2009.07.004.
[25] X. Vamvakoussi, S. Vosniadou, How many decimals are there between two fractions? aspects of secondary school students' understanding of rational numbers and their notation, Cogn. Instr. 28 (2010) 181–209, https://doi.org/10.1080/07370001003676603.
[26] M. Roell, A. Viarouge, O. Houdé, G. Borst, Inhibitory control and decimal number comparison in school-aged children, PLoS One 12 (2017) e0188276, https://doi.org/10.1371/journal.pone.0188276.
[27] A. Lubin, J. Vidal, C. Lanoë, O. Houdé, G. Borst, Inhibitory control is needed for the resolution of arithmetic word problems: a developmental negative priming study, J. Educ. Psychol. 105 (2013) 701–708, https://doi.org/10.1037/a0032625.
[28] W. De Neys, S. Rossi, O. Houdé, Bats, balls, and substitution sensitivity: cognitive misers are no happy fools, Psychon. Bull. Rev. 20 (2013) 269–273, https://doi.org/10.3758/s13423-013-0384-5.
[29] E. Travers, J.J. Rolison, A. Feeney, The time course of conflict on the Cognitive Reflection Test, Cognition 150 (2016) 109–118, https://doi.org/10.1016/j.cognition.2016.01.015.
[30] A. Miyake, N.P. Friedman, D.A. Rettinger, P. Shah, M. Hegarty, How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis, J. Exp. Psychol. Gen. 130 (2001) 621–640, https://doi.org/10.1037/0096-3445.130.4.621.
[31] W. De Neys, Automatic–heuristic and executive–analytic processing during reasoning: chronometric and dual-task considerations, Q. J. Exp. Psychol. 59 (2006) 1070–1100, https://doi.org/10.1080/02724980543000123.
[32] S. Franssens, W. De Neys, The effortless nature of conflict detection during thinking, Think. Reason. 15 (2009) 105–128, https://doi.org/10.1080/13546780802711185.
[33] M. Roell, A. Viarouge, O. Houdé, G. Borst, Inhibition of the whole number bias in decimal number comparison: a developmental negative priming study, J. Exp. Child Psychol. 177 (2019) 240–247, https://doi.org/10.1016/j.jecp.2018.08.010.
[34] N. Beauvais, A. Voudouri, E. Boissin, W. De Neys, System 2 and cognitive transparency: deliberation helps to justify sound intuitions during reasoning, Think. Reason. 32 (2025) 150–175, https://doi.org/10.1080/13546783.2025.2532653.
[35] H. Mercier, D. Sperber, Why do humans reason? Arguments for an argumentative theory, Behav. Brain Sci. 34 (2011) 57–74, https://doi.org/10.1017/S0140525X10000968.
[36] G. Pennycook, A framework for understanding reasoning errors: from fake news to climate change and beyond, in: B. Gawronski (Ed.), Adv. Exp. Soc. Psychol., Academic Press, 2023, pp. 131–208, https://doi.org/10.1016/bs.aesp.2022.11.003.
[37] W. De Neys, Defining deliberation for dual-process models of reasoning, Nat. Rev. Psychol. 4 (2025) 544–552, https://doi.org/10.1038/s44159-025-00466-6.
[38] S. Musslick, J.D. Cohen, Rationalizing constraints on the capacity for cognitive control, Trends Cogn. Sci. 25 (2021) 757–775, https://doi.org/10.1016/j.tics.2021.06.001.
[39] Z.A. Purcell, C.A. Wastell, N. Sweller, Domain-specific experience and dual-process thinking, Think. Reason. 27 (2021) 239–267, https://doi.org/10.1080/13546783.2020.1793813.
[40] M. Raoelison, V.A. Thompson, W. De Neys, The smart intuitor: cognitive capacity predicts intuitive rather than deliberate thinking, Cognition 204 (2020) 104381, https://doi.org/10.1016/j.cognition.2020.104381.
[41] E. Boissin, G. Pennycook, Who benefits from debiasing? Cognition 262 (2025) 106166 https://doi.org/10.1016/j.cognition.2025.106166.
[42] O. Ghasemi, S. Handley, S. Howarth, I.R. Newman, V.A. Thompson, Logical intuition is not really about logic, J. Exp. Psychol. Gen. 151 (2022) 2009–2028, https://doi.org/10.1037/xge0001179.
[43] O. Ghasemi, S. Handley, R. Stephens, Logical intuitions or matching heuristic? Examining the effect of deduction training on belief-based reasoning judgments, (2022). https://doi.org/10.31219/osf.io/gxd73.
[44] B.K. Hayes, R.G. Stephens, M.D. Lee, J.C. Dunn, A. Kaluve, J. Choi-Christou, N. Cruz, Always look on the bright side of logic? Testing explanations of intuitive sensitivity to logic in perceptual tasks, J. Exp. Psychol. Learn. Mem. Cogn. 48 (2022) 1598–1617, https://doi.org/10.1037/xlm0001105.
[45] C. Mekik, O. Vivier, H. Markovits, A "logical intuition" based on semantic associations, J. Exp. Psychol. Learn. Mem. Cogn. (2025), https://doi.org/10.1037/xlm0001468.
[46] C.G. Meyer-Grant, N. Cruz, H. Singmann, S. Winiger, S. Goswami, B.K. Hayes, K. C. Klauer, Are logical intuitions only make-believe? Reexamining the logic-liking effect, J. Exp. Psychol. Learn. Mem. Cogn. 49 (2023) 1280–1305, https://doi.org/10.1037/xlm0001152.

Supplementary Materials

A. Material

Study 1 : Arithmetic Word problems

The items presented here are translated into English. For the original French version, please reach out.

Practice items
Marc has 10 euros in his pocket. Marc gives 5 to his friend Thomas. How much money does he have left in his pocket? - 15 - 10 - 5 - 1
Pauline has 12 euros in her pocket. Pauline gives 6 to her friend Emilie. How much money does she have left in her pocket? - 6 - 18 - 24 - 3

Set A:

- o Conflict items:

Agathe has 30 books. She has 20 books more than Ines. How many books does Ines have? - 10 - 50 - 60 - 5
Alice has 45 flowers. She has 25 flowers more than Chloe. How many flowers does Chloe have? - 20 - 70 - 90 - 10
Antoine has 45 olives. He has 20 olives more than Victor. How many olives does Victor have? - 25 - 65 - 90 - 5
Arthur has 35 boxes. He has 5 boxes more than Pierre. How many boxes does Pierre have? - 30 - 40 - 70 - 10

Axel has 30 stamps. He has 10 stamps more than Thomas. How many stamps does Thomas have? - 20 - 40 - 60 - 10

- No-conflict items:

Lola has 40 markers. Emma has 10 markers more than Lola. How many markers does Emma have? - 50 - 25 - 75 - 5
Louis has 30 marbles. Charles has 15 marbles more than Louis. How many marbles does Charles have? - 45 - 25 - 70 - 5
Raphael has 45 coins. Maxime has 5 coins more than Raphael. How many coins does Maxime have? - 50 - 25 - 75 - 5
Lucas has 30 candies.

Lola has 40 markers. Emma has 10 markers more than Lola. How many markers does Emma have? - 50 - 25 - 75 - 5
Paul has 5 candies more than Lucas. How many candies does Paul have? - 35 - 20 - 55 - 5
Lucie has 40 pencils. Sarah has 25 pencils more than Lucie. How many pencils does Sarah have? - 65 - 35 - 100 - 5

Set B:

- Conflict items:

Elise has 50 photos. She has 15 photos more than Jade. How many photos does Jade have? - 35 - 65 - 100 - 5
Tom has 40 matches. He has 5 matches more than Alan. How many matches does Alan have? (a) 35 (b) 45 (c) 80 (d) 5
Eva has 30 bottles. She has 25 bottles more than Lea. How many bottles does Lea have? - 5 - 55 - 60 - 1
Robin has 50 newspapers. He has 20 newspapers more than Hugo. How many newspapers does Hugo have? - 30 - 70 - 100 - 10

Jules has 50 notebooks.

He has 10 notebooks more than Nathan.

How many notebooks does Nathan have?

- 40
- 60
- 100
- 20

- No-conflict items

Zoe has 35 candles. Manon has 25 candles more than Zoe. How many candles does Manon have? - 60 - 30 - 90 - 15
Julia has 40 apples. Marie has 20 apples more than Julia. How many apples does Marie have? - 60 - 30 - 90 - 15
Laura has 45 cards. Celia has 10 cards more than Laura. How many cards does Celia have? - 55 - 30 - 85 - 5
Lena has 35 plants. Julie has 10 plants more than Lena. How many plants does Julie have? - 45 - 25 - 70 - 5

Leon has 35 paintbrushes. Theo has 15 paintbrushes more than Leon. How many paintbrushes does Theo have? - 50 - 25 - 75 - 5

Study 2 : Decimal comparison items

Practice items	
Left number	**Right number**
0.2	0.789
0.975	0.83

Set A:

- o Conflict items: 1 to 10;
- o No-conflict items: 11 to 20.

Set B:

- o Conflict items: 11 to 20;
- o No-conflict items: 1 to 10.

Table S1. Decimal comparison items.

Item number	Conflict version (C)		No-conflict version (NC)	
	Left number	**Right number**	**Left number**	**Right number**
1	0.67	0.9	0.6	0.97
2	0.8	0.36	0.86	0.3
3	0.232	0.3	0.2	0.332
4	0.7	0.613	0.713	0.6
5	0.184	0.82	0.18	0.824
6	0.41	0.295	0.415	0.29
7	0.548	0.8	0.5	0.848
8	0.2	0.177	0.277	0.1
9	0.35	0.6	0.3	0.65
10	0.5	0.24	0.54	0.2
11	0.9	0.42	0.92	0.4
12	0.81	0.9	0.8	0.91
13	0.3	0.231	0.331	0.2
14	0.326	0.6	0.3	0.626
15	0.48	0.191	0.481	0.19
16	0.528	0.74	0.52	0.748
17	0.8	0.745	0.845	0.7

18	0.175	0.3	0.1	0.375
19	0.5	0.34	0.54	0.3
20	0.24	0.8	0.2	0.84

B. Response times at the final response stage

Response times were recorded at both the initial and the final response stages. We focus here on final-response times on conflict items. Means and SEM were computed on log-transformed values and back-transformed for readability.

Table S2. Mean final response times (in seconds) on conflict items, broken down by task and response accuracy. Values are means with SEM in parentheses, computed on log-transformed response times and back-transformed.

Study	Task	Correct responses	Incorrect responses
Study 1	**Arithmetic word (well-practiced task)**	6.64 (0.64)	4.94 (0.39)
	Bat-and-ball (less-practiced task)	8.45 (3.23)	4.34 (0.34)
Study 2	**Decimal comparison (well-practiced task)**	1.61 (0.10)	1.46 (0.12)
	Bat-and-ball (less-practiced task)	5.75 (1.91)	5.87 (0.51)

C. Trial exclusion analyses

Across both studies, an initial-response trial was excluded if the participant either missed the response deadline or failed the concurrent load task. To verify that these exclusions did not drive the observed task differences, we examined whether exclusion rates varied systematically across tasks and item types, and whether our conclusions held under alternative ways of handling excluded trials.

In Study 1, the proportion of excluded trials was 34.6% for arithmetic conflict items, 46.8% for arithmetic no-conflict items, 32.6% for bat-and-ball conflict items, and 25.6% for bat-and-ball no-conflict items. In Study 2, it was 37.1% for decimal conflict items, 32.4% for decimal no-conflict items, 30.4% for bat-and-ball conflict items, and 23.6% for bat-and-ball no-conflict items. We tested these differences by modeling trial exclusion (excluded vs. retained) as a function of task, fitting separate logistic mixed-effects models to conflict and no-conflict items with a by-participant random intercept. In Study 1, exclusion on the conflict items—those underlying our key analyses—did not differ between tasks (arithmetic vs. bat-and-ball: OR = 1.11, 95% CI [0.84, 1.47], p = .468); a task difference emerged only on the no-conflict control items (OR = 2.83, 95% CI [2.13, 3.77], p < .001). In Study 2, the well-practiced decimal task was excluded somewhat more than the bat-and-ball task on both conflict (OR = 1.38, 95% CI [1.09, 1.75], p = .008) and no-conflict items (OR = 1.62, 95% CI [1.25, 2.09], p < .001), reflecting the tighter 1.5-s deadline.

Because the analyses above showed that trial loss was, in some cases, higher for the well-practiced task—which could in principle inflate its retained accuracy—we further verified that our conclusions hold under two alternative treatments of the excluded trials. In each case we fitted a logistic mixed-effects model predicting accuracy on conflict items from task, with a by-participant random intercept, separately for the initial and final response stages.

First, we retained all excluded trials and coded them as incorrect (Table S2). Under this conservative coding, the well-practiced task remained substantially more accurate than the bat-and-ball task at both stages in both studies (Study 1: initial OR = 15.79, 95% CI [9.01, 27.67]; final OR = 45.44 [21.51, 96.03]; Study 2: initial OR = 50.46 [26.46, 96.22]; final OR = 129.89), all p < .001, and correct responses continued to emerge predominantly at the initial stage. For the Study 2 final-stage model, the near-floor bat-and-ball accuracy produced quasi-complete separation; the confidence interval for this odds ratio is therefore unstable and is not reported, although the effect remained highly significant (p < .001).

Second, we applied a more lenient criterion that excluded only missed-deadline trials and retained load-task failures (Table S3). The well-practiced advantage again held at both stages in both studies (Study 1: initial OR = 27.98, 95% CI [15.66, 49.99]; final OR = 65.30 [32.69, 130.41]; Study 2: initial OR = 122.84 [67.71, 222.86]; final OR = 4686.14 [1418.02, 15486.32]; all p < .001; the very large Study 2 final-stage odds ratio reflects near-floor bat-and-ball accuracy).

Together, these analyses indicate that the task differences we report are not attributable to differential trial exclusion.

Table S3. Accuracy on conflict items and direction-of-change patterns, with all excluded trials retained and coded as incorrect. Values are mean accuracy (%) for the initial and final response stages, and mean proportion (%) of each direction-of-change pattern. "11" = correct at both stages; "01" = correct only after deliberation; "10" = correct only intuitively; "00" = incorrect at both stages. Standard errors of the mean (SEM) are reported in parentheses.

Study	**Task**	**Accuracy**		**Direction of change**			
		Initial response accuracy	**Final response accuracy**	**"11" pattern**	**"01" pattern**	**"00" pattern**	**"10" pattern**

Study 1	**Arithmetic word (well-practiced task)**	29.8 (2.8)	32.6 (3.3)	14.4 (2.1)	4.8 (0.9)	77.1 (2.5)	3.7 (0.7)
	Bat-and-ball (less-practiced task)	4.0 (1.3)	3.0 (1.3)	1.9 (1.2)	1.3 (0.6)	94.8 (1.6)	2.1 (0.7)
Study 2	**Decimal comparison (well-practiced task)**	42.1 (2.7)	48.6 (3.0)	28.0 (2.0)	6.0 (0.9)	64.8 (2.3)	1.3 (0.3)
	Bat-and-ball (less-practiced task)	2.2 (1.0)	1.8 (1.0)	1.2 (0.9)	0.5 (0.3)	97.7 (1.1)	0.5 (0.9)

Table S4. Accuracy on conflict items and direction-of-change patterns under a more lenient inclusion criterion that excluded only missed-deadline trials and retained load-task failures. Values are mean accuracy (%) for the initial and final response stages, and mean proportion (%) of each direction-of-change pattern. "11" = correct at both stages; "01" = correct only after deliberation; "10" = correct only intuitively; "00" = incorrect at both stages. Standard errors of the mean (SEM) are reported in parentheses.

Study	**Task**	**Accuracy**		**Direction of change**			
		Initial response accuracy	**Final response accuracy**	**"11" pattern**	**"01" pattern**	**"00" pattern**	**"10" pattern**
Study 1	**Arithmetic word (well-practiced task)**	44.3 (3.6)	45.7 (4.1)	28.6 (3.4)	8.6 (1.7)	55.4 (4.1)	7.4 (1.6)
	Bat-and-ball (less-practiced task)	5.4 (1.6)	5.0 (1.7)	2.3 (1.4)	2.4 (0.8)	92.1 (2.0)	3.4 (1.0)
Study 2	**Decimal comparison (well-practiced task)**	64.0 (3.4)	75.7 (3.6)	54.8 (3.6)	13.1 (1.6)	29.1 (3.7)	3.0 (0.7)

Bat-and-ball (less-practiced task)	3.9 (1.4)	2.6 (1.3)	1.7 (1.1)	1.0 (0.5)	95.5 (1.6)	1.8 (0.7)

D. Alternative error strategies in decimal comparison

Following a reviewer's suggestion, we examined whether the systematic errors observed in the decimal comparison task could be attributed to heuristics other than the whole-number bias. The literature on decimal comparison has consistently identified the whole-number bias as the primary source of systematic error: students treat decimals with more digits after the decimal point as larger. We considered two alternative strategies students could in principle apply:

- Rightmost-digit strategy. Comparing the rightmost digit of each decimal and selecting the number with the larger one. For example, in 0.184 vs. 0.82, the rightmost digit of 0.184 (4) is larger than the rightmost digit of 0.82 (2), so this strategy points to 0.184 — yet 0.184 < 0.82.
- Largest-digit strategy. Comparing the largest individual digit appearing in each decimal and selecting the number containing the larger one. For example, in 0.295 vs. 0.41, the largest digit in 0.295 (9) is larger than the largest digit in 0.41 (4), so this strategy points to 0.295 — yet 0.295 < 0.41.

For each strategy, we reclassified every item pair as a conflict or a no-conflict item depending on whether applying the strategy would yield the incorrect or the correct response. Items can therefore be classified differently across strategies. The pair 0.67 vs. 0.9, for instance, is a conflict item under the whole-number bias (0.67 has more digits but is smaller), but a no-conflict item under both the rightmost-digit strategy (0.9 has the larger rightmost digit, which is the correct answer) and the largest-digit strategy (0.9 contains the larger individual digit, which is also the correct answer). We then computed mean accuracy on conflict and no-conflict items separately for each strategy at both response stages. The rationale is straightforward: if a strategy genuinely drives systematic errors, accuracy should be higher on no-conflict items—where the strategy points to the correct response—than on conflict items, where it is misleading. The size of the resulting gap can be taken as an index of how strongly a given heuristic shapes responses.

Figure S1 displays the results. Under the whole-number bias classification, accuracy was substantially lower on conflict than on no-conflict items at both stages. Under the largest-digit classification, a conflict effect was also present but markedly smaller. Under the rightmost-digit classification, no conflict effect was detectable: conflict and no-conflict accuracy were nearly identical at both stages.

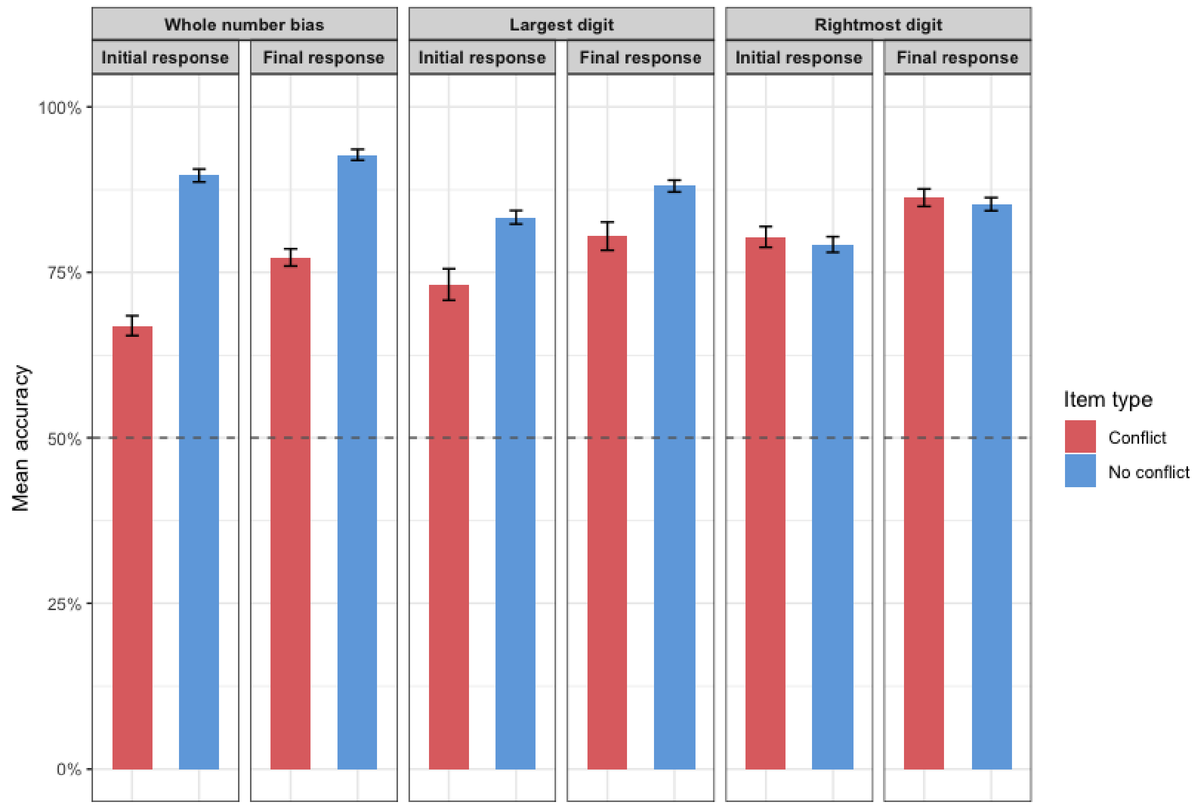

Figure S1. Mean accuracy on conflict and no-conflict items in the decimal comparison task, separately for each candidate heuristic (whole-number bias, largest-digit, rightmost-digit) and response stage (initial, final). Error bars represent standard errors of the mean (SEM).

For completeness, Figure S2 displays mean accuracy on each individual item pair, separately for the initial and final response stages, with conflict and no-conflict item accuracies according to the whole-number bias classification.

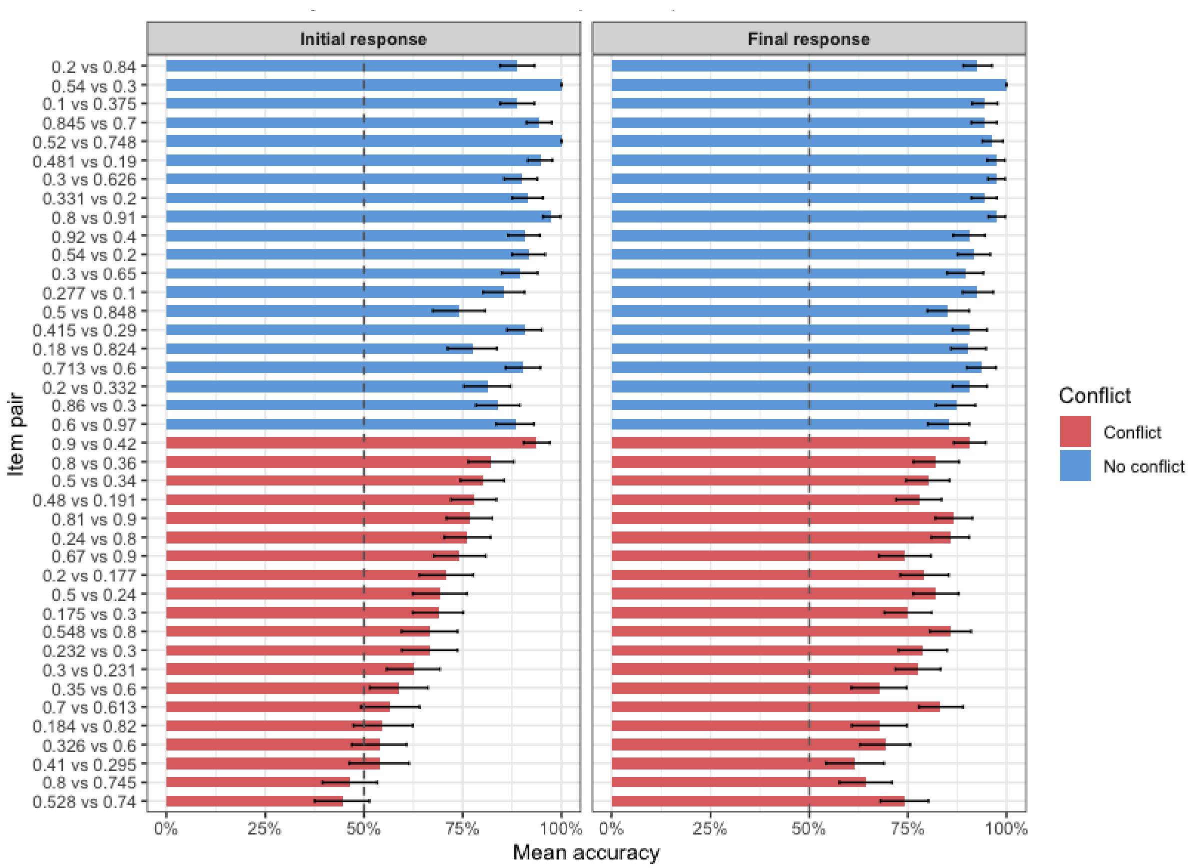

Figure S2. Mean accuracy on each item pair of the decimal comparison task, for initial and final response stages. Conflict items are pairs for which the whole-number bias points to the incorrect response; no-conflict items are pairs for which it points to the correct response. Error bars represent standard errors of the mean (SEM).