Acta Psychologica 270 (2026) 107725

Contents lists available at ScienceDirect

Acta Psychologica

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

Argument evaluation, Fast and Slow: Deliberation boosts justification strength discrimination in reasoning problems

Nicolas Beauvais [a,*], Wim De Neys [a]

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

ARTICLE INFO

Keywords:
Reasoning
Heuristics-and-biases
Argumentation
Intuition
Deliberation
Argument evaluation
Argument strength
Myside bias

ABSTRACT

While recent research shows that reasoning underpins both the production and assessment of arguments, "fast and slow" dual process theories of reasoning devoted little attention to argument evaluation. Specifically, little is known about how people distinguish between strong, explicit justifications and weaker, implicit ones, or how intuitive evaluations differ from slower, deliberate ones. Across three studies, participants solved classic reasoning problems (variants of the bat and ball problem) and rated the strength of arguments supporting different answers to these problems. Ratings were made either under time pressure and cognitive load (intuitive) or without constraints (deliberate). Results showed participants consistently gave higher ratings to stronger, explicit arguments. They also rated arguments that matched their own response generally higher than opposing arguments, under both intuitive and deliberate evaluation. Although participants showed some basic intuitive sensitivity to justification quality, deliberation significantly improved the ability to discriminate justification strength across all argument types. These findings highlight the critical role of both intuition and deliberation in argument evaluation.

1. Introduction

Decades of research showed that human reasoning is frequently undermined by biases. Popular dual process theories of reasoning attribute biases to the human tendency to rely on intuitive processes. These theories describe reasoning as an interplay between two types of processes: fast, effortless intuitive processes ("Type 1" processes) and slower, more effortful deliberative processes ("Type 2" processes) (Epstein, 1994; Evans & Stanovich, 2013; Kahneman, 2011). Although intuitive processes can be useful, they often rely on shortcuts or "heuristic" responses based on background beliefs or stereotypes that can conflict with logical principles. Consider for instance the classic "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?*" (Frederick, 2005). Most people intuitively respond that the ball costs 10 cents (Bourgeois-Gironde & Van Der Henst, 2009; Frederick, 2005) while the correct response is 5 cents: if the ball costs 5 cents, the bat will cost $1.05, and the total will be $1.10. According to traditional dual-process accounts, solving such problems correctly requires engaging in slow, deliberative reasoning to override the appealing but incorrect intuitive response (Evans & Stanovich, 2013; Kahneman, 2011; Morewedge & Kahneman, 2010).

However, a growing body of evidence suggests that if deliberation can help solve reasoning problems, it also serves other functions in the reasoning process than mere correction of intuitions (De Neys, 2025; Pennycook et al., 2015). To contrast intuitive and deliberate responses, several studies used two-response paradigms where participants must first provide a rapid, intuitive answer under time pressure to a reasoning problem, before being able to give a final response after unrestricted deliberation (Bago & De Neys, 2019; Newman et al., 2017; Raoelison et al., 2021; Thompson et al., 2011). Results showed that correct deliberative responses are generally preceded by already correct intuitive responses, suggesting that sound reasoners' intuitions often do not require correction through deliberation.

Interestingly, however, after deliberation these sound reasoners also manage to provide sound, explicit justifications for their correct responses (e.g., "the ball costs 5 cents because .05 + 1.05 = $1.10") whereas they typically fail to do so when answering intuitively –often referring to implicit processes or gut feelings (e.g., "It just felt right," "It popped into my head") (Bago & De Neys, 2019; Beauvais et al., 2026). This dual-process research suggests that deliberation also plays a key role in justifying –or rationalizing– intuitions and decisions (De Neys,

* Corresponding author at: LaPsyDE, Université Paris Cité, UMR CNRS 8240, 46 rue Saint-Jacques, 75005, Paris, France.
E-mail address: nicolas1beauvais@gmail.com (N. Beauvais).

https://doi.org/10.1016/j.actpsy.2026.107725
Received 12 January 2026; Received in revised form 25 August 2026; Accepted 26 August 2026
Available online 1 September 2026

2025; Evans, 2019; Pennycook et al., 2015).

The ability to generate justifications is crucial for communication, as it allows us to explain our opinions and decisions to others. Yet alongside the generation of justifications –and argumentation more broadly– lies the ability to assess the strength of arguments. Indeed, when people generate justifications for a position they hold, these justifications typically come with an intuitive sense of their own strength or weakness (Mercier & Sperber, 2011). Even if people are very good at rationalizing and generating justifications for whatever opinion they have, they generally produce justifications that are not nonsensical but that carry some coherence or plausibility – at least from their own perspective. Indeed, to be able to successfully convince others (if not only ourselves), producing arguments that carry at least some degree of coherence or plausibility is a functional requirement, which presupposes some sensitivity to argument strength. Likewise, when others present arguments to persuade us, the capacity to assess their strength is essential to avoid being misled (Mercier, 2021; Mercier & Sperber, 2011). Effective argument evaluation is thus important both to construct convincing justifications for one's position and to critically assess the arguments of others.

While the close connection between reasoning and argumentation is now well established (Hornikx & Hahn, 2012; Mercier & Sperber, 2011, 2017; Svedholm-Häkkinen & Kiikeri, 2022), argument evaluation received little attention within the dual process theory framework. In particular, it remains unclear whether individuals can effectively distinguish different types of arguments, or how they judge different justifications offered for specific claims. For example, when solving reasoning problems, are people able to recognize the strength of a well-justified, explicit argument (e.g., "the ball costs 5 cents because.05 + 1.05 = $1.10") over a more implicit one referring to intuition (e.g., "I just felt it")?

In this paper, we aim to explore how reasoners evaluate such different types of justifications. While explicit justifications may be viewed as more valuable, as they allow for external verification and assessment, this assumption has not been directly tested. Indeed, cultural narratives and popular media–from Malcolm Gladwell's *Blink* (2007) to public figures like Steve Jobs (Isaacson, 2013)– often celebrate intuitive processing, portraying it as a mark of brilliance. Moreover, empirical studies show that people treat intuition as an appropriate, sometimes prescribed, basis for certain decisions (Inbar et al., 2010; Oktar & Lombrozo, 2022), report benefits from affective-focused, gut-based choice (Mikels et al., 2011), experience such choices as more expressive of their true self and hold them with greater certainty (Maglio & Reich, 2019), and may even follow intuitions after recognizing them to be unjustified (Walco & Risen, 2017). Hence, in theory, people might judge justifications referring to more implicit, intuition-based reasoning as stronger or more convincing. We aim to address this question by directly comparing how people evaluate arguments that vary in justification explicitness.

Interestingly, work from the field of informal argumentation suggests that individuals do seem to have a fairly good capacity to distinguish strong arguments from weak ones (Corner & Hahn, 2009; Hahn & Oaksford, 2007; Hoeken et al., 2014; Marin et al., 2024; Svedholm-Häkkinen et al., 2025). For example, arguments violating normative criteria such as ad hominem arguments or other fallacies are generally perceived as weaker and less convincing than arguments that better satisfy standards of argument quality (Deans-Browne & Singmann, 2026; Hoeken et al., 2014; Svedholm-Häkkinen & Kiikeri, 2022). However, the focus of these studies lay on informal arguments and not on the classic, formal reasoning problems central to dual process theories, in which prior beliefs cue a strong biasing heuristic response that conflicts with the normatively correct response (as in the bat-and-ball problem). Nevertheless, this body of work lends credence to the idea that reasoners may have some competency in distinguishing different types of arguments, including the kinds of justifications offered for responses to formal reasoning problems. Note that while formal reasoning tasks like the bat-and-ball problem have mathematically correct solutions –contrary to informal logic tasks where arguments concern debatable claims with no objectively correct answer– people still produce and defend different answers to such problems, with more or less explicit justifications, and benefit from discussing them with other individuals (Laughlin & Ellis, 1986; Trouche et al., 2014). As such, although the process by which people solve such problems is mathematical in nature, these problems are also concerned with issues of production and evaluation of arguments (used here in the broad sense of a connected series of statements used to establish a definite proposition, see e.g. Cook et al., 2018). For instance, when presented with a tentative answer –especially someone else's–a reasoner must appraise how compelling the justification supporting it is. This evaluative, downstream process is conceptually distinct from solving, and it is to this process that the argument-evaluation literature directly speaks. While developed in a different context, this literature provides relevant insights into how people assess the quality of reasons offered in support of a conclusion, and we therefore draw on it here as a broader theoretical backdrop.

More generally, Mercier and Sperber's argumentative theory of reasoning (2009, 2011, 2017) also entails that the human mind may be especially well equipped to discern argument quality. Critically, these authors indicate that this capacity may be mainly intuitive. However, this potential intuitive dimension of argument evaluation received little direct testing (Hornikx et al., 2022). Some evidence suggests that instructing participants to think analytically –rather than intuitively– does not improve their ability to detect fallacies (Svedholm-Häkkinen & Kiikeri, 2022) or evaluate informal arguments (Svedholm-Häkkinen & Hietanen, 2024). Hornikx et al. (2022) on the other hand found mixed results when testing how participants judge informal argument strength by forcing people to rely on intuitive processing with a time pressure design. In sum, more work is needed to understand the respective contribution of intuitive and deliberate judgments in argument evaluation. Moreover, the limited evidence available comes almost entirely from informal argumentation, leaving open whether the same holds for the evaluation of justifications in formal reasoning problems, where a normatively correct answer competes with an appealing intuitive one. Hence, in the present study, in addition to establishing whether individuals can effectively distinguish different types of arguments in a classic formal reasoning task (Study 1) we also wanted to establish the nature of the underlying process –whether it relies on fast, intuitive or slow, more deliberate thinking (Studies 2–3).

Across our three studies, participants first solved bat-and-ball problems before evaluating arguments supporting different answers to these problems. Arguments varied systematically in both response correctness and justification explicitness (i.e., a more implicit reference to intuition – "I just felt it" or a more explicit reference specifying actual calculations). Recording participants' responses to the problem enabled us to directly examine how evaluations differed depending on participants' ability to correctly solve the problem. For example, whether biased participants who themselves failed to solve the problem nevertheless showed sensitivity to (sound) argument explicitness.

Study 1 tested whether sound reasoners who gave correct responses to the problems and biased reasoners who gave incorrect responses show sensitivity to the different argument types. In Studies 2 and 3, half of the participants rated arguments under cognitive load and time pressure –conditions prompting intuitive evaluations– while the other half evaluated arguments without constraints, allowing for deliberation (Thompson et al., 2011), to test the nature of the argument evaluation process.

2. Study 1

2.1. Methods

2.1.1. Pre-registration and data availability

All studies were preregistered on the Open Science Framework. The preregistrations, data, material and analyses scripts for all studies are available on OSF at https://osf.io/qz7ph/.

2.1.2. Participants

We recruited 102 participants on Prolific Academic (www.prolific.ac). All participants provided informed consent before participation and were compensated according to Prolific Academic's guidelines. We included only native English speakers from the United Kingdom, United States, Ireland, Australia, Canada, or New Zealand. Participants' mean age was 39.25 years (SD = 13.26), and 49% identified as female. Regarding education level, 75.5% had a university degree and 23.5% a high school diploma. Only one participant reported an education level of less than high school.

2.1.3. Material

2.1.3.1. Problems. Participants were presented with six problems, adapted from Frederick's (2005) classical "bat-and-ball" problem. To minimize potential familiarity bias (Chandler et al., 2014; Haigh, 2016; Stieger & Reips, 2016), we used structurally similar but content-altered versions of the original problem (Bago & De Neys, 2019; Beauvais et al., 2026; Raoelison & De Neys, 2019). Each problem presented participants with the total count of two items and described the specific relationship between their quantities. Participants needed to determine the quantity of one item based on this information, choosing from four answer options. An example format is as follows:

In a building residents have 370 dogs and cats in total.
There are 300 more dogs than cats.
How many cats are there?

- *7*
- *35*
- *70*
- *105*

In each problem, participants selected one of four answer options: the correct answer (i.e., "5 cents" for the classic bat-and-ball problem, "35" in the previous example), the intuitive "heuristic" response ("10 cents" in the classic problem, "70" in the example), and two foil choices which were always the sum of the correct and heuristics options ("15 cents" in the original bat-and-ball units, "105" in the example) and their second greatest common divider ("1 cent" in original units, "7" in the example). The correct answer option was derived using the form of the mathematical equation for solving the standard bat-and-ball problem: $100 + 2\times = 110$. The heuristic answer option was obtained by using the equation thought to lead to the erroneous initial intuitive response in these problems: $100 + x = 110$ (Bago & De Neys, 2019; Kahneman, 2011). For each problem, the presentation order of the four response options was randomized.

Half of the problems (three out of six) were presented in a "conflict" format, where the intuitive heuristic response diverged from the correct answer, as in the classic bat-and-ball problem. The remaining three problems were control "no-conflict" problems, where the logically correct answer and heuristic reasoning aligned. This alignment was achieved by omitting the critical relational statement "more than" (De Neys et al., 2013). The no-conflict version of the previous example presented the question as follows:

In a building, residents have 370 dogs and cats in total.
There are 300 dogs.
How many cats are there?

These no-conflict trials allowed to verify whether participants remained attentive throughout the study, as near-perfect accuracy would be expected without random guessing.

2.1.3.2. Arguments. Following each problem response, participants were presented with seven distinct arguments supporting possible answers to the problem they had just seen (labelled A-G, see Table 1). Each argument consisted of a proposed answer accompanied by a justification (i.e., for the previous example: "*There are [X] cats, because [Y]*", where X is the proposed answer and Y its justification). We presented different types of justifications (see below) based on the type of justification that participants generated for similar problems in prior studies (Bago & De Neys, 2019; Beauvais et al., 2026).

The proposed answer was either the correct answer, the heuristic answer, or a non-heuristic incorrect answer. The justification was either implicit, referring to an unspecified intuition (e.g., "because it readily came to mind"), or explicit, referring to a specific calculation which could be valid or invalid. Explicit valid justifications provided a correct mathematical rationale for the proposed answer, whereas explicit invalid justifications presented syntactically plausible but mathematically irrelevant reasoning that happened to yield the proposed answer. The latter allowed us to check that participants did not rate arguments positively merely because they contained a calculation, without assessing its relevance to the problem.

Hence, for each problem we constructed three correct-response arguments that differed in their justification type: an implicit justification (e.g., in original bat-and-ball units, "because the response that readily comes to mind is 5", argument A in Table 1); an explicit valid justification giving the correct rationale (e.g., "because if the bat costs \$1 more and the total is \$1.10, the bat must cost \$1.05 and the ball 5 cents", argument B); and an explicit invalid justification (e.g., "because 50 cents minus 45 is 5 cents, so the solution must be that the ball costs 5 cents", argument C). Likewise, we constructed three heuristic-response arguments differing in justification type: an implicit justification (argument D); an explicit justification that is mathematically valid *for the heuristic answer* but does not yield the correct response (e.g., "because subtracting

Table 1
Argument types.

Answer type	Justification type	Example	Label
Correct (5 cents)	Implicit	"The ball costs 5 cents, because the response that readily comes to mind when considering the problem is 5."	A
	Explicit valid	"The ball costs 5 cents, because if the bat costs \$1 more and the total is \$1.10, the bat must cost \$1.05 and the ball 5 cents."	B
	Explicit invalid	"The ball costs 5 cents, because 50 cents minus 45 is 5 cents, so the solution must be that the ball costs 5 cents."	C
Incorrect heuristic (10 cents)	Implicit	"The ball costs 10 cents, because the response that readily comes to mind when considering the problem is 10."	D
	Explicit heuristic	"The ball costs 10 cents, because subtracting \$1 to the total of \$1.10 amounts to a total of 10 cents."	E
	Explicit invalid	"The ball costs 10 cents, because if the bat costs \$1, then minus 50 is 50 cents, and 50 cents minus 40 amounts to 10 cents."	F
Incorrect control (15 cents)	Control invalid	"The ball costs 15 cents, because \$1 more minus 90 is 10 cents, and adding 5 amounts to a total of 15 cents."	G

\$1 from the total of \$1.10 amounts to a total of 10 cents", argument E)–which, to avoid confusion with the explicit valid justification for the correct answer, we refer to as the "explicit heuristic" justification; and an explicit invalid justification (e.g., "because if the ball costs \$1, then minus 50 is 50 cents, and 50 cents minus 40 amounts to 10 cents", argument F). Finally, the last argument with the non-heuristic incorrect answer ("15 cents" in original bat-and-ball units) was always paired with an explicit invalid justification (e.g., "because \$1 more minus 90 is 10 cents, and adding 5 amounts to a total of 15 cents", argument G); for simplicity, we refer to it as the "control invalid" argument. The seven argument types are summarized in Table 1, with examples adapted to the classic bat-and-ball problem; the actual arguments used for each problem are provided in Supplementary material A.

Each argument was controlled in length and phrasing to prevent participants from using these features as heuristics for evaluation. Arguments averaged 22 words in length, with variation within each problem limited to ±3 words. To ensure justification structure did not signal correctness, features present in explicit valid justifications of correct response arguments were included in other justifications: "if […] then" statements were included in explicit invalid justifications of incorrect heuristic response arguments, and the critical word "more" appeared in explicit invalid justifications of control invalid arguments. Justifications were standardized across problems, with minor adjustments only for trial-specific terms (e.g., items like dogs or cats). In no-conflict problems, the justifications of correct arguments corresponded to the justifications of incorrect heuristic arguments in conflict problems. Correspondingly, the justifications of incorrect arguments in no-conflict problems perfectly mirrored the justifications of correct arguments in conflict problems.

2.1.4. Procedure

The experiment was run online on Qualtrics (www.qualtrics.com). Participants began the study by reading a brief introduction that outlined that their task would be to answer reasoning problems and evaluate arguments. Participants were familiarized with the task structure with two practice trials, before engaging with the six problems and their corresponding arguments in the actual experiment. The content of the practice trials differed from the actual problems and arguments to prevent contamination of the main data. For each problem trial, sentences were presented serially on the screen, with each sentence displayed for 2 s. Upon viewing the complete problem, participants were asked to select one of the four response options, and their response times were recorded. After they selected a response, the problem text remained visible at the top of the screen, and participants proceeded to evaluate the presented arguments. Arguments were shown one at a time, and the presentation order of arguments was randomized for each problem. Participants were asked to assess each argument's strength on a 0 ("very weak") to 10 ("very strong") Likert scale (following recommendation from Leung, 2011; Tanujaya et al., 2022; Wu & Leung, 2017), with the question: "*How strong do you think this argument is?*" displayed above the scale. Participants could take all the time they wanted to submit their evaluation. Once all arguments for a problem had been evaluated, a new problem was displayed. The presentation order of the problems was randomized. Upon completion of the experiment, participants were presented with the standard bat-and-ball problem and asked whether they had encountered it before. They were then instructed to provide the solution in a free-response format. Lastly, participants completed demographic questions (age, gender, education level) and were presented with a debriefing message.

2.1.5. Exclusion criteria

At the end of the experiment, 33 participants reported having seen, solved, or read about the classical bat-and-ball problem before and responded correctly to this problem. According to the preregistration, we performed separate data analyses with and without participants who were familiar with the classic version of the bat-and-ball problem. Both analyses yielded similar results. Hence, as planned in the preregistration we report the results of the full analysis including all participants. Results of the control analyses with these participants excluded can be found in the Supplementary Material D.

2.2. Results

2.2.1. Accuracy

Consistent with the literature, participants generally struggled to correctly solve conflict bat-and-ball problems. On average, accuracy on these problems reached 32.68% (SEM = 4.31). Among incorrect responses, 99.03% corresponded to the heuristic answer. As expected, participants' performance on control no-conflict problems, where the heuristic answer is the correct one, was at ceiling (M = 99.35%, SEM = 0.46).

2.2.2. Argument evaluation

2.2.2.1. Conflict problems. We compared the mean ratings attributed to each argument based on participants' own response to the problem (correct or incorrect) as well as the argument's answer type and justification. We explored the results for conflict trials on which participants solved the problem correctly ($N = 39$, 100 correct responses) and incorrectly ($N = 78$, 206 incorrect responses) separately. Results are presented in Fig. 1. In each of these groups, we also analyzed the correct (left of the dashed line) and incorrect response arguments (right of the dashed line) separately.

Fig. 1 (left panel) shows that reasoners who solved the problem correctly gave very high ratings to explicit valid justifications for the correct answer (M = 8.95). Bonferroni-corrected contrast analyses indicated these were significantly higher than ratings for implicit, (M = 2.25, $t(38) = 14.57$, $p < .001$, $d = 2.33$) and explicit invalid (M = 1.92, $t(38) = 14.87$, $p < .001$, $d = 2.38$) justifications. Hence, perhaps not surprisingly, sound reasoners consistently rated valid explicit arguments higher than invalid ones but critically, they also found implicit justifications relatively unpersuasive. Note that these reasoners gave similar ratings to implicit and explicit invalid justifications for correct response arguments ($t(38) = 0.39$, $p = .696$, $d = 0.13$).

As Fig. 1 (left panel) further indicates, reasoners who solved the problem correctly gave overall very low ratings to arguments supporting the incorrect response. A repeated-measure ANOVA confirmed these ratings were overall significantly lower than those for correct-response arguments, $F(3.80, 144.56) = 118.65$, $p < .001$, $\eta^2_G = 0.641$. Nevertheless, here too contrast analyses indicated these reasoners rated explicit heuristic arguments (M = 2.11) higher than implicit ones (M = 1.03, $t(38) = 3.87$, $p = .002$, $d = 0.62$), as well as higher than both explicit-invalid and control-invalid arguments (all $3.1 < t < 3.5$, $p < .05$, $0.5 < d < 0.6$). Implicit arguments received similar ratings than explicit invalid (M = 1.25, $t(38) = 0.22$, $p = .561$, $d = 0.21$) and control invalid (M = 0.94, $t(38) = 0.07$, $p = .990$, $d = 0.05$) arguments.

Turning to the trials on which participants answered the conflict problem incorrectly, Fig. 1 (right panel) shows that here participants gave the highest ratings to explicit heuristic justifications for the incorrect heuristic answer (i.e., "10 cents because 110 minus 100 is 10") (M = 8.39). Bonferroni-corrected contrast analyses showed that these were significantly higher than ratings for implicit (M = 4.60, $t(77) = 9.24$, $p <. 001$, $d = 1.05$), or any other justifications for the incorrect answers ($M_{explicit\text{-}invalid} = 3.29$, $t(77) = 12.62$, $p < .001$, $d = 1.43$; $M_{control\text{-}invalid} = 0.70$, $t(77) = 20.83$, $p < .001$, $d = 2.36$). Implicit justifications for the incorrect heuristic answer received higher ratings than explicit invalid justifications ($t(77) = 2.94$, $p = .022$, $d = 0.33$), and than control invalid ones ($t(77) = 8.18$, $p < .001$, $d = 0.93$).

Furthermore, strength ratings for correct response arguments (left of dashed line in Fig. 1, right panel) were markedly lower than those for incorrect response arguments, as confirmed by a repeated-measures

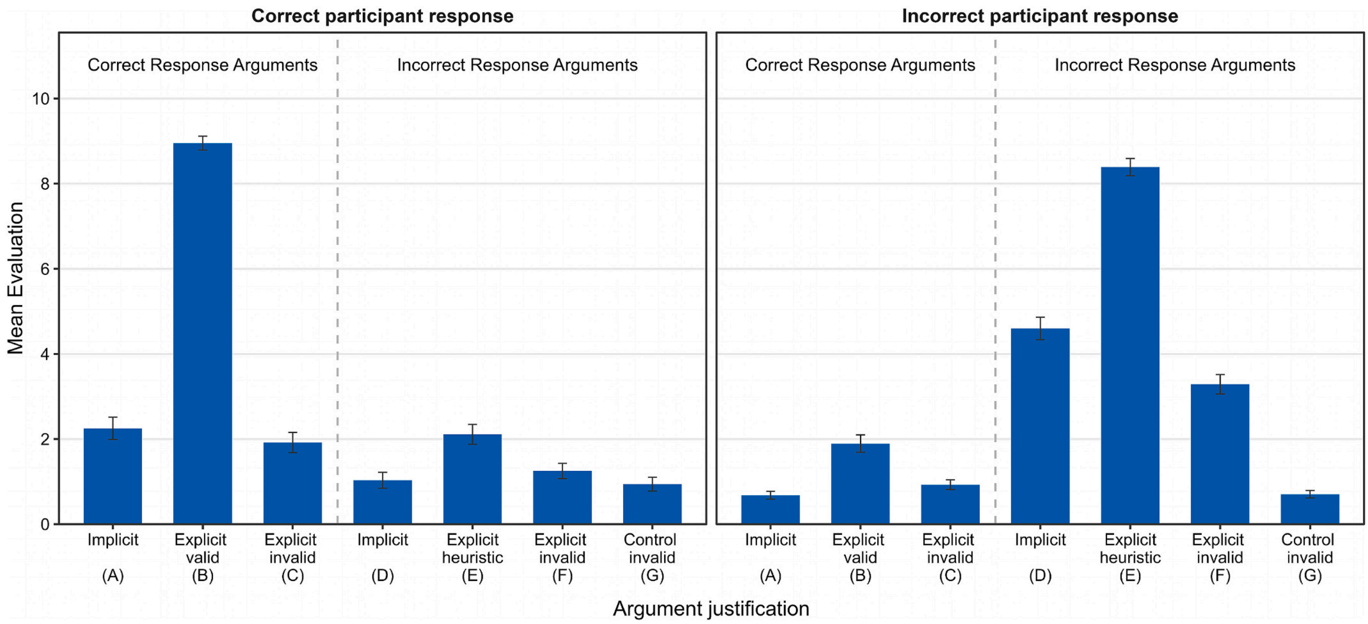

Fig. 1. Mean evaluations of arguments on conflict problems after correct (left panel) and incorrect (right panel) participant response, by argument response and justification (Study 1). Error bars are standard errors of the mean.

ANOVA, $F(2.98, 229.05) = 86.25$, $p < .001$, $\eta^2_G = 0.454$. Nevertheless, even among correct response arguments, these participants gave slightly higher ratings to explicit valid justifications (M = 1.89) than to explicit invalid (M = 0.93, $t(77) = 4.80$, $p < .001$, $d = 0.61$), or implicit ones (M = 0.68, $t(77) = 5.34$, $p < .001$, $d = 0.54$). They gave similar ratings to implicit and explicit invalid justifications for the correct response arguments ($t(77) = 0.37$, $p = .116$, $d = 0.23$). In sum, even though these participants failed to solve the problem correctly and consistently gave higher ratings to arguments supporting the heuristic incorrect response, their argument evaluations showed some minimal sensitivity to the correct response logic justification.

2.2.2.2. No-conflict problems. For completeness, we also analyzed results for the no-conflict (control) problems. The interested reader can find a full overview of the results in Supplementary Material C. Note that here too, explicit valid justifications overall received higher ratings than both implicit and explicit invalid arguments.

2.2.2.3. Same argument contrasts. In an additional analysis, we examined whether participants adjusted their evaluation of an explicit valid argument depending on the normative context. As noted earlier, explicit heuristic arguments for the incorrect heuristic response in conflict problems were identical in wording to explicit valid arguments for the correct response in no-conflict problems (i.e., in classic bat-and-ball units they would both use the justification "10 cents because 1.10 minus 1 equals 10 cents"). While the content of the argument was identical, its normative status differed: in no-conflict problems the argument was appropriate, while in conflict problems it was erroneous. To isolate sensitivity to this distinction, we compared evaluations from reasoners who correctly solved no-conflict problems with those from reasoners who gave incorrect heuristic responses to conflict problems. In both cases, the argument thus supported the participant's answer, allowing us to control for alignment in testing ratings' sensitivity to the normative context. Results showed that the argument was rated significantly higher when it was normatively appropriate (M = 8.97) than when it was not (M = 8.21, $t(77) = 2.80$, $p = .006$, $d = 0.32$).

We also examined a second case in which the same argument was in different normative contexts: we compared explicit valid arguments for the correct response in conflict problems to explicit valid arguments for the incorrect response in no-conflict problems (in classic bat-and-ball units, these arguments would support the "5 cents" answer with a valid justification). These arguments were thus normatively appropriate in conflict problems but erroneous in no-conflict problems. We again contrasted ratings from reasoners who gave incorrect heuristic answers on conflict problems with those from reasoners who correctly solved no-conflict problems. In both cases the argument contradicted the participant's answer, but was appropriate in conflict problems and not in no-conflict ones, again allowing to test evaluations' sensitivity to the appropriateness of the argument beyond alignment with participants' response. Results showed that participants rated the arguments significantly higher when it was appropriate (M = 1.89) than when it was erroneous (M = 0.97, $t(77) = 4.75$, $p < .001$, $d = 0.54$). Together, these findings suggest that even participants who were biased and responded heuristically when solving a conflict problem nevertheless showed some sensitivity to a justification's normative status when evaluating arguments.

3. Studies 2 and 3

Study 1 indicated that reasoners can effectively distinguish different types of arguments. They judged explicit arguments as stronger than implicit ones and demonstrated some basic sensitivity to an argument's normative status –even when they failed to solve the problem correctly. In addition, participants generally rated arguments supporting their own response higher than opposing arguments. In Study 2 and 3 we tested the nature of the underlying process –whether it relies on fast, intuitive or slow, more deliberate thinking. To contrast intuitive and deliberative evaluations, we manipulated cognitive constraints during argument evaluations in Study 2. Half of the participants evaluated arguments under a strict time limit and cognitive load, while the other half evaluated arguments without constraints. Study 3 was a replication of Study 2 with stricter constraints. Given that the methods and results of these studies were highly similar, we present them alongside each other.

3.1. Methods

3.1.1. Pre-registration and data availability

All studies were preregistered on the Open Science Framework. The preregistrations, data, material and analyses scripts for these studies are available on OSF at https://osf.io/qz7ph/.

3.1.2. Participants

In study 2, we recruited 258 participants on Prolific Academic (www.prolific.ac). All participants provided informed consent before participation and were compensated according to Prolific Academic's guidelines. We included only native English speakers from the United Kingdom, United States, Ireland, Australia, Canada, or New Zealand. Participants' mean age was 37.52 years (SD = 12.92), 49.6% were female. Regarding education level, 64.7% had a university degree and 34.5% a high school diploma. Two participants reported an education level of less than high school.

In study 3, we recruited 261 participants on Prolific Academic (www.prolific.ac), recruited from the same countries as previously listed. All participants provided informed consent before participation and were compensated according to Prolific Academic's guidelines. Participants' mean age was 37.93 years (SD = 12.8), 49.4% were female. Regarding education level, 62.8% had a university degree and 36.8% a high school diploma. Only one participant reported an education level of less than high school.

3.1.3. Material

The same bat-and-ball problems and corresponding arguments from Study 1 were used in Studies 2 and 3. However, in both these studies we used a “fast” vs. “slow” between-subjects design where half of the participants evaluated each argument under a strict time limit and cognitive load task (fast, intuitive condition), while the other half of participants evaluated arguments without constraints (slow, deliberate condition). The combination of time limit and concurrent cognitive load task has been successfully used in reasoning tasks to burden executive resources in order to minimize the possibility of deliberation being involved in the response production, allowing to capture an “intuitive” answer (Bago & De Neys, 2017, 2019; Boissin et al., 2021; Franiatte et al., 2024; Franssens & De Neys, 2009; Miyake et al., 2001; Raoelison & De Neys, 2019; Thompson et al., 2011). Note that in the fast-intuitive condition of Studies 2 and 3, participants faced no time constraints when responding to the problems themselves and could take as much time as needed to submit their responses. The time limit and cognitive load only applied to the argument evaluations. Instructions made this clear for the participants, and an additional reminder about it was included at the end of the practice session in Study 3.

The load task was a memorization task in which participants had to retain a complex visual pattern presented for 2 s before the argument. In Study 2, the pattern was a 3 × 3 grid in which 4 crosses were placed. In Study 3 the pattern was a 4 × 4 grid with 5 crosses (see Supplementary Material A for examples). In both studies, after rating the argument participants were presented with four different matrices and asked to identify the pattern presented before the argument (see Bago & De Neys, 2017 for further details). Feedback was provided to indicate whether their choice was correct or not.

In addition to the cognitive load task, participants in the fast-intuitive condition had to rate each argument under a strict time limit. In Study 2, this time limit was set at 5 s (determined by rounding the first quartile of response times from Study 1 to the nearest whole number). To signal the approaching deadline, the screen background changed to yellow 1.5 s before the time limit. In Study 3, the time limit was reduced to 4 s to further minimize the possibility of deliberation, and the screen background turned yellow 750 milliseconds before the deadline.

3.1.4. Procedure

Both studies were conducted online using Qualtrics (www.qualtrics.com). Participants began by providing informed consent and reading a brief introduction outlining their tasks, which included solving reasoning problems, evaluating arguments, and performing an additional task in the fast-intuitive condition. Participants completed two practice trials to familiarize themselves with the experimental tasks. These trials introduced the problem solving and argument evaluation tasks in both conditions and included the cognitive load and time limit constraints in the fast-intuitive condition.

Participants were presented with the same set of problems and arguments as in Study 1. The display and response format of problems and arguments in the slow-deliberate condition of Studies 2 and 3 were the same as in Study 1. In the fast-intuitive condition, the problem-solving format was similar to Study 1, but argument evaluations incorporated the time constraint and cognitive load. Before each argument, a visual matrix was displayed for 2 s as part of the cognitive load task. Participants then viewed the argument and had 5 s (Study 2) or 4 s (Study 3) to rate its strength on the 0–10 Likert scale. After rating the argument, participants were shown four visual matrices and asked to identify the one they had memorized. Feedback was provided on their choice before the presentation of the next set of matrix and argument. The presentation order of problems and arguments was randomized. At the end of both studies, participants completed demographic questions (age, gender, education level) and were presented with a debriefing message.

3.1.5. Exclusion criteria

As in Study 1, at the end of Studies 2 and 3 participants were presented with the standard bat-and-ball problem, asked if they had encountered it before, and instructed to provide the solution in a free-response format. There were 52 participants who were familiar with the problem and responded correctly to it in Study 2, 61 in Study 3. For both studies, the analyses with and without exclusion of these participants gave similar results. Consequently, as specified in the preregistration, we report the results of the full analysis. Results of the control analyses with these participants excluded can be found in the Supplementary Material D.

In the fast-intuitive condition of Studies 2 and 3, argument evaluation trials were excluded from the analysis if participants did not respond before the time limit or failed to correctly recall the matrix presented before the trial. In Study 2, participants missed the deadline in 2.99% of trials and failed to recall the correct matrix in 15.0% of trials, leaving 82.0% of trials for analysis. In Study 3, participants in the fast-intuitive condition missed the deadline in 5.94% of trials and failed to recall the correct matrix in 29.8% of trials, leaving 64.3% of trials for analysis. No trials were excluded from the slow-deliberate condition in both studies.

3.2. Results

3.2.1. Accuracy

Across Studies 2 and 3, participants consistently showed poor performance on conflict problems, in line with Study 1 and the existing literature. Mean accuracy for participants of the slow-deliberate condition was 36.34% (SEM = 3.92) in Study 2, and 32.61% (SEM = 3.72) in Study 3. Surprisingly, despite identical problem response conditions in both fast-intuitive and slow-deliberate conditions (i.e., the constraints were only applied during the argument evaluation phase not during the problem response phase) participants from the fast-intuitive condition had a lower accuracy on conflict problems, with only 16.0% (SEM = 2.87) of correct answers in Study 2. This difference was significant under a between-subject ANOVA, $F(1, 256) = 17.15$, $p < .001$, $\eta^2_G = 0.063$. This pattern persisted in Study 3: mean accuracy for participants of the fast-intuitive condition was 15.45% (SEM = 3.04), significantly lower than for participants of the slow-deliberate condition, $F(1, 259) = 12.42$, $p < .001$, $\eta^2_G = 0.046$. Taken together, these analyses suggest that participants from the fast-intuitive condition were slightly less accurate in solving the problems than participants from the slow-deliberate condition, probably because constraints on argument evaluations influenced their approach to solving problems and induced quicker responses. This was confirmed by latency analyses which pointed to faster responses among participants from the fast-intuitive condition compared to responses from participants of the slow-deliberate condition, in both studies (see Supplementary material B).

Note that heuristic-incorrect answers accounted for 98.10% and 99.61% of all incorrect responses in the fast-intuitive and slow-deliberate conditions of Study 2, respectively, and for 98.08% and 98.92% of all incorrect responses in the corresponding conditions of Study 3.

3.2.2. Argument evaluation

3.2.2.1. Conflict problems. In Studies 2 and 3, we contrasted fast-intuitive evaluations of arguments with slow-deliberate ones. Fig. 2 shows the results for both Study 2 (Fig. 2A) and Study 3 (Fig. 2B). We again analyzed the results for conflict trials on which participants solved the problem correctly (Study 2: Fast-intuitive condition: N = 30, 60 correct responses, Slow-deliberate condition: N = 56, 145 correct responses; Study 3: Fast-intuitive condition: N = 23, 57 correct responses, Slow-deliberate condition: N = 53, 135 correct responses) and incorrectly (Study 2: Fast-intuitive condition: N = 113, 315 incorrect responses, Slow-deliberate condition: N = 94, 254 incorrect responses; Study 3: Fast-intuitive condition: N = 110, 312 incorrect responses, Slow-deliberate condition: N = 104, 279 incorrect responses) separately. In each of these groups, we also analyzed the correct (left of the dashed line) and incorrect response arguments (right of dashed line) separately. We tested the effect of evaluation condition (fast-intuitive vs. slow-deliberate) on the argument types in question with mixed-design (repeated-measures) ANOVAs.

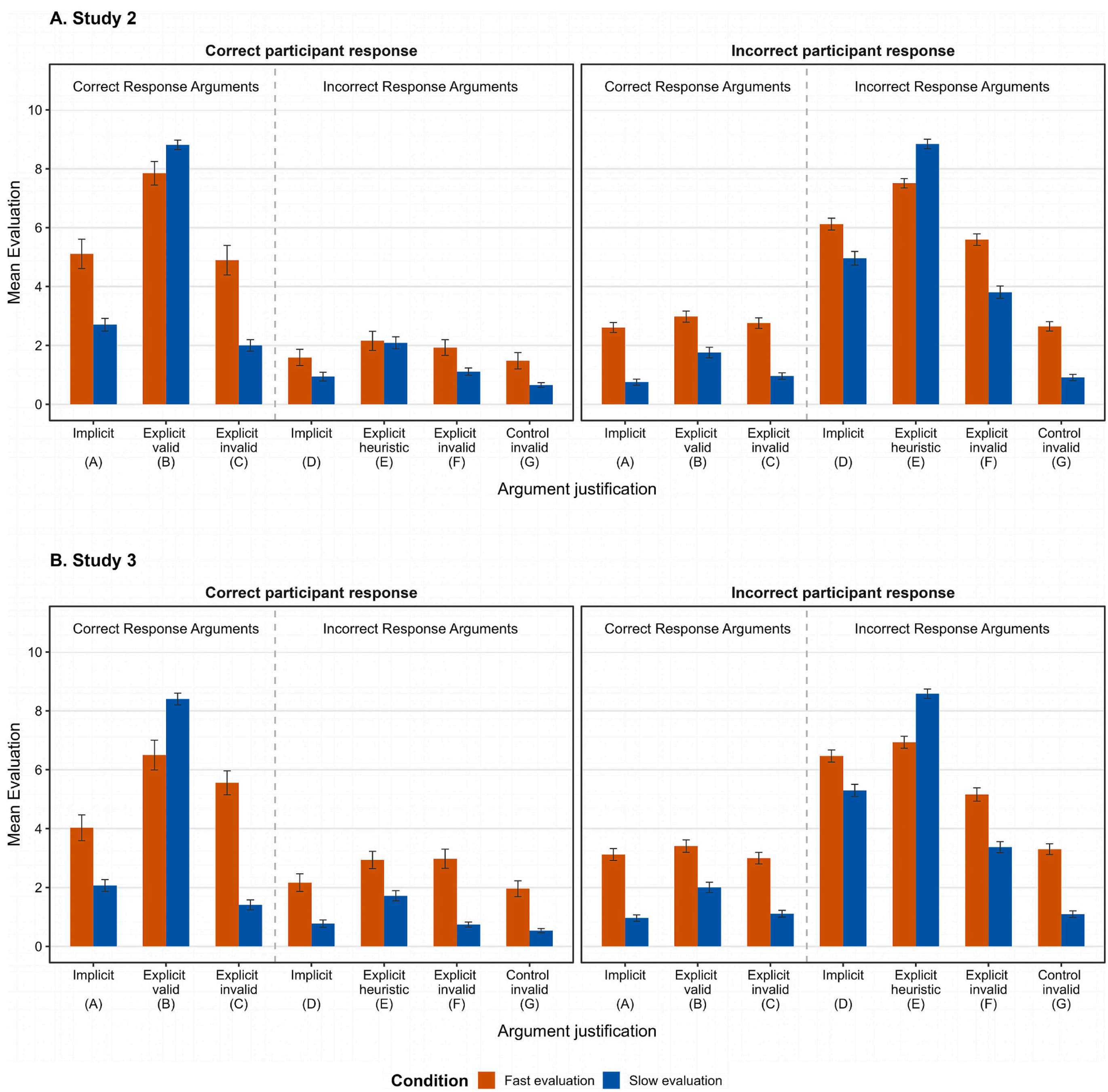

Fig. 2. Mean evaluations of arguments on conflict problems by condition in Study 2 (A) and Study 3 (B), separated by participant response to the problem and argument response correctness. Error bars are standard errors of the mean.

Starting with conflict trials on which participants solved the problem correctly, Fig. 2 shows that for arguments supporting the correct response (left of the dashed line), in the slow trials of both Study 2 (N = 145 trials, for all arguments) and 3 (N = 135 trials, for all arguments) we replicate the pattern observed in Study 1 with higher strength ratings for explicit valid arguments (Study 2: M = 8.81, Study 3: M = 8.40), compared to implicit (Study 2: M = 2.70, $t(79) = 15.21$, $p < .001$, d = 2.32; Study 3: M = 2.07, $t(66) = 15.64$, $p < .001$, $d = 2.10$) and explicit invalid arguments (Study 2: M = 2.00, $t(79) = 19.28$, $p < .001$, $d = 2.99$; Study 3: M = 1.41, $t(66) = 19.06$, $p < .001$, $d = 2.41$). As Fig. 2 further indicates, in the fast-intuitive condition, this pattern of stronger ratings for explicit valid arguments was also observed but became less pronounced. Repeated-measures ANOVAs on correct-response arguments among correct responders revealed a significant interaction between condition (fast-intuitive vs. slow-deliberate) and argument in both Study 2, $F(1.96, 154.79) = 15.10$, $p < .001$, $\eta^2_G = 0.084$, and Study 3, $F(1.77, 116.48) = 29.83$, $p < .001$, $\eta^2_G = 0.150$. Bonferroni-corrected contrast analyses showed implicit and explicit invalid arguments received significantly higher ratings under fast-intuitive than under slow-deliberate evaluations (Study 2: all $3.4 < t < 4.6$, $p < .001$, $0.76 < d < 1.21$; Study 3: all $2.0 < t < 5.0$, $p < .05$, $0.70 < d < 1.7$), while correct explicit valid arguments tended to receive slightly lower ratings under fast-intuitive than under slow-deliberate evaluations (Study 2: $t(79) = -1.78$, $p = .080$, $d = -0.49$; Study 3: $t(66) = -3.53$, $p < .001$, $d = -1.10$). Interestingly, Bonferroni-corrected tests indicated that the pattern of higher ratings for explicit valid arguments (Study 2: M = 7.95, N = 46 trials; Study 3: M = 6.50, N = 32 trials) over implicit ones remained significant in the fast-intuitive condition (Study 2: $M_{implicit}$ = 5.11, N = 54 trials, $t(79) = 4.99$, $p < .001$, $d = 0.85$; Study 3: $M_{implicit}$ = 4.03, N = 35 trials, $t(66) = 3.20$, $p = .006$, $d = 0.90$). The pattern of higher ratings for valid over invalid explicit arguments in the fast-intuitive condition was also significant in Study 2 ($M_{explicit\text{-}invalid}$ = 4.90, N = 48 trials, $t(79) = 6.64$, $p < .001$, $d = 0.90$), but only tendential in Study 3 ($M_{explicit\text{-}invalid}$ = 5.56, N = 45 trials, $t(66) = 1.76$, $p = .247$, d = 0.27). Note that in both Study 2 and Study 3, in the two conditions the ratings of implicit arguments did not significantly differ from those of explicit invalid arguments (all $p > .05$, $0.21 < \Delta M < 1.53$). Overall, these results indicate that the pattern of higher ratings for explicit valid arguments is still observed under intuitive processing conditions but is attenuated.

For correct reasoners' evaluation of arguments supporting the incorrect response (left panels of Fig. 2A and B, right of the dashed line), a similar pattern was observed in Study 2 and 3. Bonferroni-corrected contrast analyses showed that in the slow-deliberate condition, although these arguments received overall low ratings, those with explicit heuristic justifications received slightly higher ratings (Study 2: M = 2.09, N = 145 trials; Study 3: M = 1.72, N = 135 trials) than those with implicit or invalid justifications (Study 2: all $4.8 < t < 5.7$, $p < .001$, $0.63 < d < 0.73$; Study 3: all $4.1 < t < 5.6$, $p < .001$, $0.52 < d < 0.71$; see Supplementary Material C for full details), just as we observed in Study 1. However, in the fast-intuitive condition, the pattern of higher ratings for explicit arguments was less clear. Repeated-measures ANOVAs on incorrect-response arguments among correct responders revealed a non-significant interaction between condition and argument in Study 2, $F(2.60, 212.89) = 2.45$, $p = .074$, $\eta^2_G = 0.010$, but a significant interaction in Study 3, $F(2.28, 169.17) = 3.32$, $p = .032$, $\eta^2_G = 0.016$. Bonferroni-corrected contrast analyses revealed no difference between fast-intuitive and slow-deliberate evaluations of heuristic arguments in Study 2 (all $0.4 < t < 1.9$, $p > .05$, $0.02 < d < 0.42$), but indicated that all heuristic arguments received significantly higher ratings under fast-intuitive than under slow-deliberate evaluations in Study 3 (all $2.3 < t < 6.2$, $p < .05$, $0.66 < d < 1.46$). Under fast-intuitive evaluation, reasoners who correctly solved the problem gave similar ratings to all incorrect-response arguments in Study 2, $F(2.66, 71.87) = 1.84$, p = .199, $\eta^2_G = 0.017$. In Study 3, contrast analyses showed these reasoners intuitively rated explicit heuristic (M = 2.93, N = 45 trials) and explicit invalid (M = 2.98, N = 42 trials) arguments similarly ($p = 1.0$), but both slightly higher than implicit (M = 2.17, N = 42 trials) and control invalid (M = 1.96, N = 49 trials) arguments (all $2.7 < t < 5.4$, $p < .05$, $0.54 < d < 0.77$). Hence, correct responders showed very little tendency to judge explicit heuristic justifications as stronger when rating incorrect-response arguments intuitively. As a sidenote, across the remaining conditions in Study 2 and 3, the ratings of implicit arguments did not significantly differ from those of explicit invalid arguments (all $p > .05$, $0.03 < \Delta M < 0.33$, see Supplementary material C for full specifications; the only significant difference, in the fast-intuitive condition of Study 3, is reported above).

Note that as in Study 1, in the slow-deliberate condition correct responders gave overall higher ratings to correct-response arguments than to incorrect-response arguments, as confirmed by repeated-measures ANOVAs: Study 2: $F(1, 55) = 236.69$, $p < .001$, $\eta^2_G = 0.547$, Study 3: $F(1, 52) = 237.85$, $p < .001$, $\eta^2_G = 0.582$. This difference was not affected by constraints, Study 2: $F(1, 84) = 2.87$, $p = .094$, $\eta^2_G = 0.012$, Study 3: $F(1, 74) = 1.28$, $p = .261$, $\eta^2_G = 0.006$. In the fast-intuitive condition –just as in the slow-deliberate one– correct responders rated correct-response arguments overall higher than incorrect-response ones, Study 2: $F(1, 29) = 75.36$, $p < .001$, $\eta^2_G = 0.554$; Study 3: $F(1,22) = 26.32$, $p < .001$, $\eta^2_G = 0.304$.

Turning to the results for reasoners who gave heuristic incorrect responses, Fig. 2 shows that for arguments supporting the incorrect response (right panels of Fig. 2A and B, right of the dashed line), the slow trials of both Study 2 and 3 replicate the pattern observed in Study 1, with clear higher ratings for explicit heuristic justifications (Study 2: M = 8.84; Study 3: 8.58) over any other justification (Study 2: all $11.6 < t < 23.5$, $p < .001$, $1.08 < d < 2.49$, N = 254 trials; Study 3: all $11.2 < t < 19.4$, $p < .001$, $0.94 < d < 2.01$, N = 279 trials; see Supplementary Material C for the full specification). As Fig. 2 further illustrates, in the fast-intuitive condition, this pattern of higher ratings for explicit heuristic arguments persisted but was less pronounced. Repeated-measures ANOVAs on incorrect-response arguments among incorrect responders revealed significant interactions between condition and argument in both Study 2, $F(2.75, 542.41) = 16.80$, $p < .001$, $\eta^2_G = 0.047$, and Study 3, $F(2.26, 176) = 40.59$, $p < .001$, $\eta^2_G = 0.231$. Bonferroni-corrected contrast analyses showed explicit heuristic arguments received lower ratings under fast-intuitive evaluation than under slow-deliberate evaluation (Study 2: $t(197) = -3.19$, $p = .002$, $d = -0.51$; Study 3: $t(181) = -3.13$, $p = .002$, $d = -0.48$), whereas all other arguments received higher ratings under intuitive evaluation (Study 2: all $2.9 < t < 6.1$, $p < .01$, $0.39 < d < 0.90$; Study 3: all $3.2 < t < 6.7$, $p < .001$, $0.50 < d < 0.94$). Interestingly, Bonferroni-corrected tests indicated that the pattern of higher ratings for explicit heuristic justifications (Study 2: M = 7.51, N = 268 trials; Study 3: M = 6.93, N = 193 trials) over explicit invalid ones remained significant in the intuitive condition (Study 2: $M_{explicit\text{-}invalid}$ = 5.60, N = 219 trials, $t(197) = 7.08$, $p < .001$, $d = 0.71$; Study 3: $M_{explicit\text{-}invalid}$ = 5.16, N = 197 trials, $t(181) = 5.06$, $p < .001$, d = 0.63). The pattern of higher ratings for explicit heuristic over implicit arguments was significant in Study 2 ($M_{implicit}$ = 6.12, N = 275 trials, $t(197) = 4.58$, $p < .001$, $d = 0.47$) but not in Study 3 ($M_{implicit}$ = 6.47, N = 201 trials, $t(181) = 1.22$, $p = 1.0$, $d = 0.22$). Here again, results suggest that if the pattern of higher ratings for explicit heuristic arguments remains present under intuitive processing conditions, it is attenuated compared to the deliberate condition.

The interested reader will note that in the two conditions of both Study 2 and Study 3, incorrect responders almost always rated implicit incorrect-response arguments higher than explicit invalid ones. The former arguments received significantly higher ratings than the latter in all but the fast-intuitive condition of Study 2 ($p = .636$, $\Delta M = 0.52$; all other $3.54 < t < 7.65$; $p < .01$; $0.32 < d < 0.87$).

Concerning incorrect reasoners' evaluation of correct-response arguments (right panels of Fig. 2A and B, left of the dashed line), similar patterns emerged in Studies 2 and 3. In the slow-deliberate condition, Bonferroni-corrected contrast analyses showed that explicit valid justifications (Study 2: M $= 1.76$; Study 3: M $= 2.00$) received slightly higher ratings than implicit or invalid ones (Study 2: all $N = 254$ trials, $5.31 < t < 5.33$, $p < .001$, $0.45 < d < 0.48$; Study 3: all $N = 279$ trials, $6.0 < t < 6.5$, $p < .001$, $0.51 < d < 0.55$; see Supplementary material C for full details), again as we found in Study 1. However, in the fast-intuitive condition, the pattern of higher ratings for explicit arguments was less clear. Repeated-measures ANOVAs on correct-response arguments among incorrect responders revealed significant interactions between condition and argument in both Study 2, $F(1.81, 355.44) = 6.64$, $p < .01$, $\eta^2_G = 0.008$, and Study 3, $F(1.70, 328.30) = 8.34$, $p < .001$, $\eta^2_G = 0.013$. Bonferroni-corrected contrast analyses showed all correct-response arguments received significantly higher ratings under fast-intuitive than under slow-deliberate evaluations in Study 2 (all $2.0 < t < 5.7$, $p < .05$, $0.30 < d < 0.85$), and all arguments ($5.4 < t < 6.3$, $p < .001$, $0.81 < d < 0.95$) but explicit valid ones ($t(193) = 1.53$, $p = .128$, $d = 0.23$) did so in Study 3. Under fast-intuitive evaluation, reasoners who incorrectly solved the problem rated all correct-response arguments alike in both Study 2, $F(1.91, 197.02) = 2.56$, $p = .083$, $\eta^2_G = 0.017$, and Study 3, $F(1.98, 178.14) = 2.34$, $p = .10$, $\eta^2_G = 0.005$. Hence, incorrect responders did not judge explicit valid justifications as stronger when rating correct-response arguments intuitively. Note that in both Study 2 and Study 3, these reasoners rated implicit and explicit invalid correct arguments similarly, in both conditions (all $p > .05$, $0.13 < \Delta M < 0.21$).

Thus, when arguments were evaluated intuitively, both correct and incorrect responders showed diminished sensitivity to differences in argument strength. Distinctions based on justification type–that is, between strong explicit valid arguments and implicit or invalid arguments–became less clear.

In addition, note that here again in the slow-deliberate condition incorrect responders gave overall higher ratings to incorrect-response arguments than to correct-response arguments, as confirmed by repeated-measures ANOVAs: Study 2: $F(1, 93) = 204.98$, $p < .001$, $\eta^2_G = 0.532$, Study 3: $F(1,103) = 138.20$, $p < .001$, $\eta^2_G = 0.455$. This difference was slightly affected by condition in Study 2: $F(1, 205) = 4.40$, $p = .037$, $\eta^2_G = 0.008$, but not in Study 3: $F(1,211) = 2.51$, $p = .115$, $\eta^2_G = 0.005$. In the fast-intuitive condition –just as in the slow-deliberate one– incorrect responders rated incorrect-response arguments overall higher than correct-response ones, Study 2: $F(1,112) = 134.79$, $p < .001$, $\eta^2_G = 0.36$; Study 3: $F(1, 108) = 90.89$, $p < .001$, $\eta^2_G = 0.294$. Notably, this preference for response-aligned arguments was strong enough that incorrect responders rated implicit arguments supporting their own (incorrect) answer significantly higher than explicit valid arguments supporting the correct answer, under both fast-intuitive evaluation (Study 2: M $= 6.12$ vs. 2.98, $t(190) = 7.28$, $p < .001$, $d = 0.51$; Study 3: M $= 6.47$ vs. 3.41, $t(173) = 5.43$, $p < .001$, $d = 0.37$) and slow-deliberate evaluation (Study 2: M $= 4.96$ vs. 1.76, $t(190) = 5.91$, $p < .001$, $d = 0.41$; Study 3: M $= 5.30$ vs. 2.00, $t(173) = 4.85$, $p < .001$, $d = 0.33$). Thus, alignment with one's own answer could outweigh the explicitness of a justification when judging argument strength. This specific pattern was, however, largely absent among correct responders: they rated aligned implicit arguments higher than opposing explicit heuristic arguments only in the fast-intuitive condition of Study 2 (M $= 5.11$ vs. 2.16, $t(77) = 3.62$, $p = .011$, $d = 0.39$), with no significant difference in the slow-deliberate condition or in either condition of Study 3 (all $p > .05$).

3.2.2.2. No-conflict problems. For completeness, we also analyzed results for the no-conflict (control) problems. The interested reader can find a full overview of the results in Supplementary Material C. Here again, as accuracy rates were at ceiling, we did not analyze the few trials on which participants responded incorrectly. Results from the deliberate condition were identical to that of Study 1, with explicit valid justifications receiving higher ratings than both implicit and explicit invalid arguments. As with conflict problems in Studies 2–3, these differences were attenuated under intuitive conditions, reflecting a reduced sensitivity to justification strength.

3.2.2.3. Same argument contrasts. We conducted the same-argument contrast analysis for both intuitive and deliberate judgments from Studies 2 and 3. As a reminder, this analysis tested whether participants adjusted their evaluation of an explicit valid argument depending on the context in which it was presented. We focus first on the "10 cents" arguments with an explicit heuristic justification (i.e., in classic bat-and-ball units, "10 cents because 1.10 minus 1 is 10 cents"). While argument content remained identical from conflict to no-conflict problems, normative status changed: in no-conflict problems this argument was appropriate and valid, while in conflict problems it was erroneous. We compared the ratings attributed to these arguments by incorrect responders on conflict problems with those from correct responders on no-conflict problems. In both cases the argument aligned with the participant response, allowing to test for the rating difference while controlling for alignment. Across Studies 2 and 3, we tested whether the difference varied depending on the evaluation condition. Fig. 3 (left panel) presents the mean absolute difference by condition in each study. In Study 2, although this difference appears descriptively larger in the slow-deliberate condition than in the fast-intuitive one, a mixed ANOVA revealed no significant interaction, $F(1,201) = 1.51$, $p = .221$, $\eta^2_G = 0.001$. In Study 3, on the other hand, the interaction was significant, $F(1, 194) = 8.74$, $p = .004$, $\eta^2_G = 0.007$. Post hoc tests showed a robust effect in the slow-deliberate condition where the argument was rated higher when actually appropriate (M $= 9.16$) than when not (M $= 8.18$, $t(194) = 4.83$, $p < .001$, $d = 0.40$), while ratings did not significantly differ in the fast-intuitive condition, $t(194) = 0.49$, $p = .628$, $d = 0.07$.

We also compared the ratings attributed to correct explicit valid arguments on conflict problems to those attributed to incorrect explicit valid arguments on no-conflict problems. These arguments correspond to "5 cents" arguments with explicit valid justifications (i.e., in classic bat-and-ball units, "5 cents because .05 plus 1.05 equals \$1.10"). They were thus normatively appropriate in conflict problems but erroneous in no-conflict problems. We contrasted ratings from reasoners who gave incorrect heuristic answers on conflict problems with those from reasoners who correctly solved no-conflict problems. This allowed to control for alignment with participant response when testing whether ratings differed by context (in both cases, the argument contradicted the participant's answer). We tested whether this difference depended on the evaluation condition in Studies 2 and 3. Fig. 3 (right panel) presents the mean absolute difference by condition in each study. Mixed ANOVAs showed significant interactions between condition and context in both Study 2, $F(1, 196) = 5.95$, $p = .016$, $\eta^2_G = 0.006$, and Study 3, $F(1, 190) = 10.81$, $p = .001$, $\eta^2_G = 0.015$. Post hoc comparisons confirmed a significant effect in the slow-deliberate condition, with participants rating the argument higher when it was appropriate (Study 2: M $= 2.17$; Study 3: M $= 2.69$) than when it was erroneous in both Study 2 (M $= 1.07$, $t(196) = 4.97$, $p < .001$, $d = 0.47$) and Study 3 (M $= 1.09$, $t(190) = 6.41$, $p < .001$, $d = 0.53$). No reliable differences were found in the fast-intuitive condition, Study 2: $t(196) = 1.72$, $p = .087$, $d = 0.19$; Study 3: $t(190) = 1.43$, $p = .155$, $d = 0.22$.

Taken together, these results suggest that participants were mostly sensitive to the normative appropriateness of arguments when reasoning deliberately. Under intuitive evaluation, there were no –or very little– difference in ratings between cases where the argument was normatively appropriate and cases where it was erroneous.

4. General discussion

The present research investigated how people evaluate arguments in reasoning problems and how sensitivity to argument strength depends on whether evaluations are made intuitively or deliberately. In Study 1, participants rated arguments with various types of justification without constraints. Results consistently showed that reasoners can effectively distinguish different types of arguments. They gave higher strength ratings to explicit valid (or heuristic) arguments matching their own response; yet, even among arguments supporting the opposing response,

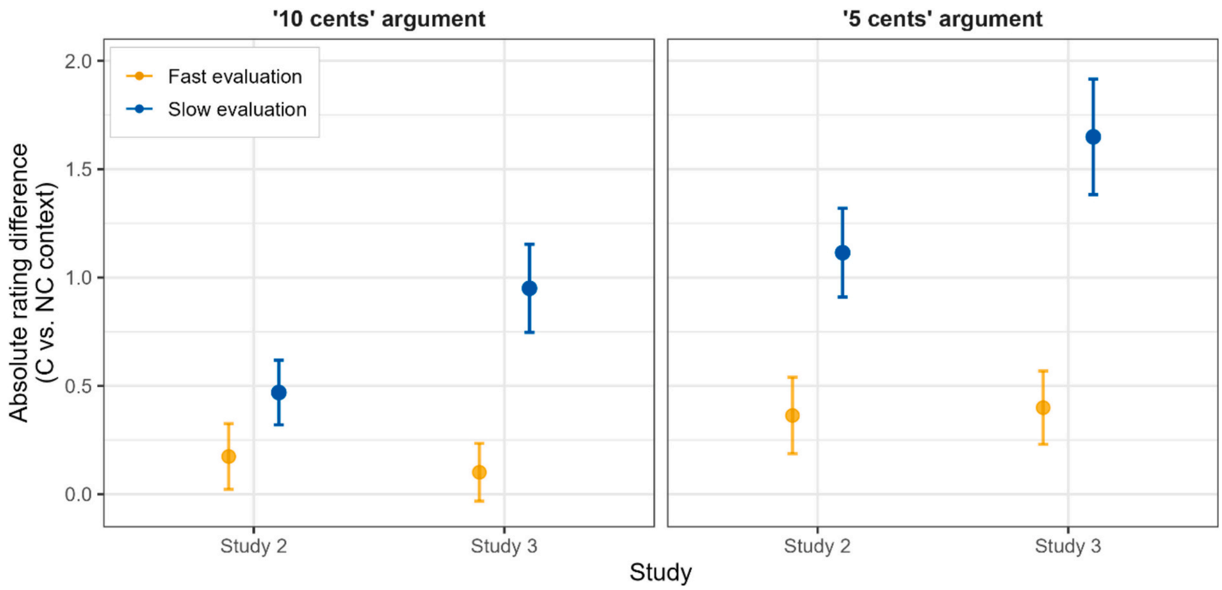

Fig. 3. Absolute differences in mean ratings between conflict (C) and no-conflict (NC) contexts for the "10 cents" and "5 cents" explicit arguments, plotted separately for fast-intuitive and slow-deliberate evaluations in Studies 2 and 3. Error bars are standard error of the mean. Differences for fast-intuitive evaluations are shown with reduced opacity to reflect the absence of significant effects.

they also rated those with explicit justifications higher than the implicit or invalid ones. They further demonstrated some basic sensitivity to an argument's normative status, even when they failed to solve the problem correctly.

In Studies 2 and 3, we examined the nature of the underlying process –whether it relies on fast, intuitive or slow, more deliberate thinking– by manipulating cognitive constraints during argument evaluation. In the slow-deliberate (unconstrained) condition, results replicated the patterns observed in Study 1: participants gave higher ratings to explicit valid arguments and demonstrated sensitivity to normative status, confirming the robustness of these effects. In the fast-intuitive (constrained) condition, the pattern of higher ratings for explicit valid arguments persisted but was weaker than in the slow-deliberate condition. In addition, sensitivity to normative status was negligible: arguments were not rated significantly higher when appropriate, unlike in the slow-deliberate condition.

Taken together, in line with Mercier and Sperber (2009, 2011), our results indicate that sensitivity to argument strength (here in reasoning problems) appears to be at least partly intuitive: participants generally judged explicit valid justifications as stronger even when evaluating arguments under cognitive constraints. However, it is also clearly enhanced by deliberation. Reasoners who had time to reflect when judging arguments made sharper distinctions between arguments that differed in justification type and showed sensitivity to an argument's normative status, indicating that deliberation enhances sensitivity to argument strength beyond what intuition allows. This supports the view that engaging in deliberate, analytic thinking improves argument assessment, consistent with recent findings on analytic thinking and information evaluation skills (e.g., Bago et al., 2020; Pennycook & Rand, 2019; Sultan et al., 2024; Svedholm-Häkkinen et al., 2024).

Note that perhaps unsurprisingly, participants tended to rate arguments that aligned with their own response (i.e., correct-response arguments for correct responders, and in over 98% of cases, incorrect heuristic-response arguments for incorrect responders) overall more favorably than arguments opposing it. These results echo prior findings from the informal argumentation literature that showed individuals tend to judge arguments that align with their prior beliefs or opinions more positively (Deans-Browne & Singmann, 2026; Edwards & Smith, 1996; Lord et al., 1979; Taber & Lodge, 2006; Wolfe & Kurby, 2017), a phenomenon often referred to as myside bias (Baron, 1995, p. 199; Mercier, 2022; Stanovich et al., 2013). Interestingly, in our data, this effect occurred both under intuitive and deliberate evaluations. This influence of alignment was at times strong enough to outweigh the explicitness and soundness of a justification: among reasoners who solved the problem incorrectly, implicit arguments supporting their own (incorrect) answer were judged stronger than explicit, valid arguments supporting the correct answer, even under deliberation. Notably, this was largely specific to incorrect responders, suggesting that these reasoners may be more prone to myside bias.

Furthermore, looking at alignment in more detail, results showed that if sensitivity to argument strength was generally lower in intuitive than in deliberate evaluations, it was particularly reduced when arguments contradicted participants' own responses. In the fast-intuitive conditions, reasoners who correctly answered conflict problems rated all incorrect-response arguments uniformly low, and those who gave incorrect answers rated correct-response arguments almost indistinctively low. In contrast, when evaluating arguments that aligned with their answers, these reasoners favored well-justified arguments over weaker ones even in intuitive evaluations, as previously mentioned. Participants thus tended to pay less attention to justification strength when intuitively judging opposing arguments than when considering aligning arguments. In the slow-deliberate conditions, however, participants gave higher ratings to explicit valid and explicit heuristic arguments both when they aligned with their own response and when they opposed it. In other words, sensitivity to justification strength was present for arguments aligned with participants' own response under both intuitive and deliberate evaluation; what deliberation added was the extension of this sensitivity to opposing arguments –as well as to an argument's normative status. Intuitive evaluation thus depended more heavily on response alignment, whereas deliberation supported a more systematic appraisal of justification strength, sharpening sensitivity across argument types.

This role of deliberation in argument evaluation, here in the context of classic reasoning problems, is also consistent with recent evolutions in dual process theories of reasoning (De Neys, 2025; Evans, 2019; Pennycook, 2022; Pennycook et al., 2015), which highlighted that deliberation can serve multiple purpose beyond those traditionally associated with it such as response generation and correction. In particular, several studies highlighted the role played by deliberation in response justification (Bago & De Neys, 2019; Beauvais et al., 2026). The present findings show that deliberation helps not only in the production of justifications (or arguments) for one's decisions but also in the evaluation of arguments proposed by others, boosting argument strength discrimination ability. This suggests that deliberation plays a crucial role in monitoring and assessing the quality of information and reasons beyond generating one's own responses. In doing so, it contributes to more effective reasoning in dialogical and social contexts, where judging the strength of others' arguments is as important as forming one's own.

In addition, this set of studies also provides some evidence on the perceived strength of implicit arguments. Although intuitive processing is often praised – from popular books (Gladwell, 2007) to Steve Jobs attributing his success to following his gut feeling (Isaacson, 2013) – and is positively valued by individuals for certain decisions (Inbar et al., 2010; Maglio & Reich, 2019; Mikels et al., 2011; Oktar & Lombrozo, 2022), implicit arguments that used references to intuitions as justifications generally received lower ratings in our studies than arguments for the same answer that used explicit valid (or heuristic) justifications. This was true for all deliberate judgments, but also for a number of intuitive ones. Note that implicit arguments generally received similar ratings as explicit invalid arguments. However, interestingly, incorrect responders in the deliberate conditions rated incorrect heuristic-response arguments with implicit justifications higher than those with explicit invalid justifications, suggesting that these participants were slightly more lenient toward implicit justifications for arguments that aligned with their own responses.

Clearly, our work also faces some limitations. First, we only looked at one type of formal reasoning problem (bat-and-ball problems). Although it is one of the most widely adopted tasks in the dual process field (Bago & De Neys, 2019; Boissin et al., 2021; Bourgeois-Gironde & Van Der Henst, 2009; Franiatte et al., 2024; Meyer & Frederick, 2022; Raoelison & De Neys, 2019), future research should explore whether the findings generalize to other reasoning tasks.

Second, although consistency in response accuracy across the different conflict problems was rather high, a given participant did not always respond consistently across these problems, sometimes contributing both correct and incorrect trials. Our correct- and incorrect-response analyses are therefore conducted at the trial level, and the two sets of trials are not drawn from two fully distinct groups of participants. The resulting data are thus noisier than a fully between-participants comparison would yield, and caution is thus warranted in interpreting them as reflecting two stable groups of reasoners. To address this, we conducted an additional analysis restricting the comparison to participants who responded consistently correctly or consistently incorrectly across all conflict trials; the resulting pattern closely replicated that of the main trial-level analysis in all three studies (see Supplementary material E), suggesting this limitation does not meaningfully impact our conclusions.

Third, although we minimized deliberation using strict, validated procedures, one can never be completely certain that processing in constraint conditions was purely intuitive (De Neys, 2023). Nevertheless, the results show that argument evaluation can be maintained to some degree even when deliberation is severely hampered.

Fourth, an alternative reading of the attenuation observed under time pressure and load is that these constraints increased random error rather than selectively reducing sensitivity to argument strength. We cannot entirely distinguish these accounts with the present data: ratings were generally compressed toward the scale mean under constraints, and because the distinctions between argument types under deliberate evaluation differed in size to begin with (e.g., were larger between aligning than between opposing arguments), a uniform increase in noise could render smaller distinctions non-significant while preserving larger ones. We therefore remain cautious about interpreting the reduced sensitivity under constraints as a purely qualitative change in processing. Importantly, however, this ambiguity does not affect our central conclusion: whether the weaker discrimination under intuitive evaluation reflects genuinely reduced sensitivity through qualitatively distinct processing or noisier responding, deliberation was consistently associated with sharper discrimination of justification strength. Additionally, the advantage of explicit, well-justified arguments aligned with participants' own response remained robust even under constraints, indicating that some sensitivity to justification strength was retained when deliberation was minimized. This pattern replicated across Studies 2 and 3 –holding under even stricter constraints in Study 3– attesting to its robustness and arguing against the possibility that judgments under constraint were purely random.

Lastly, previous work has indicated that individuals also show some degree of sensitivity to argument strength in informal, every day argumentation (Corner & Hahn, 2009; Deans-Browne & Singmann, 2026; Hahn & Oaksford, 2007; Hoeken et al., 2014; Marin et al., 2024; Svedholm-Häkkinen et al., 2025). Generalizing the current constraints findings to such informal, everyday arguments would be an interesting follow-up avenue.

In conclusion, the present studies demonstrated that individuals can meaningfully discriminate the strength of justifications offered for responses to reasoning problems, consistently judging explicit, well-justified arguments as stronger. They also shed light on the respective contributions of intuitive and deliberate reasoning to argument evaluation: while both intuitive and deliberate evaluations were influenced by response alignment, intuitive judgments already supported the pattern of higher strength ratings for explicit, well-justified arguments, and deliberation further enhanced the ability to discriminate justification strength and to recognize normative appropriateness.

CRediT authorship contribution statement

Nicolas Beauvais: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wim De Neys:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Ethics approval statement

This research was approved by the ethical committee from the Université Paris Cité (CER U-Paris Cité, N° IRB: 00012023-146).

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used GPT-4 to assist in proofreading the manuscript for grammatical and stylistic improvements. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgments

Preparation of this manuscript benefitted from support from the European Union under Horizon Europe Programme Grant Agreement no. 101120763 – TANGO. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2026.107725.

Data availability

Links to the data, code and material for all studies have been included in the manuscript at the "Pre-registration and data availability" subsections.

References

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition, 158*, 90–109. https://doi.org/10.1016/j.cognition.2016.10.014

Bago, B., & De Neys, W. (2019). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking and Reasoning, 25*(3), 257–299. https://doi.org/10.1080/13546783.2018.1507949

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General, 149*(8), 1608–1613. https://doi.org/10.1037/xge0000729

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking & Reasoning, 1*(3), 221–235. https://doi.org/10.1080/13546789508256909

Beauvais, N., Voudouri, A., Boissin, E., & De Neys, W. (2026). System 2 and cognitive transparency: Deliberation helps to justify sound intuitions during reasoning. *Thinking & Reasoning, 32*(1), 150–175. https://doi.org/10.1080/13546783.2025.2532653

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

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

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112–130. https://doi.org/10.3758/s13428-013-0365-7

Cook, J., Ellerton, P., & Kinkead, D. (2018). Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters, 13*(2), Article 024018. https://doi.org/10.1088/1748-9326/aaa49f

Corner, A., & Hahn, U. (2009). Evaluating science arguments: Evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied, 15*(3), 199–212. https://doi.org/10.1037/a0016533

De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences, 46*, 1–71. https://doi.org/10.1017/S0140525X2200142X

De Neys, W. (2025). Defining deliberation for dual-process models of reasoning. *Nature Reviews Psychology, 4*(8), 544–552. https://doi.org/10.1038/s44159-025-00466-6

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

Deans-Browne, C., & Singmann, H. (2026). For everyday arguments prior beliefs play a larger role on perceived argument quality than argument quality itself. *Cognition, 266*, Article 106257. https://doi.org/10.1016/j.cognition.2025.106257

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology, 71*(1), 5–24. https://doi.org/10.1037/0022-3514.71.1.5

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist, 49*(8), 709–724.

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

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

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

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

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

Gladwell, M. (2007). *Blink: The power of thinking without thinking; [By the author of The Tipping Point]* (1. Back Bay trade pbk. ed.). Back Bay Books.

Hahn, U., & Oaksford, M. (2007). The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review, 114*(3), 704–732. https://doi.org/10.1037/0033-295X.114.3.704

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

Hoeken, H., Šorm, E., & Schellens, P. J. (2014). Arguing about the likelihood of consequences: Laypeople's criteria to distinguish strong arguments from weak ones. *Thinking and Reasoning, 20*(1), 77–98. https://doi.org/10.1080/13546783.2013.807303

Hornikx, J., & Hahn, U. (2012). Reasoning and argumentation: Towards an integrated psychology of argumentation. *Thinking & Reasoning, 18*(3), 225–243. https://doi.org/10.1080/13546783.2012.674715

Hornikx, J., Weerman, A., & Hoeken, H. (2022). An exploratory test of an intuitive evaluation method of perceived argument strength. *Studies in Communication Sciences, 22*(2), Article 2. https://doi.org/10.24434/j.scoms.2022.02.003

Inbar, Y., Cone, J., & Gilovich, T. (2010). People's intuitions about intuitive insight and intuitive choice. *Journal of Personality and Social Psychology, 99*(2), 232–247. https://doi.org/10.1037/a0020215

Isaacson, W. (2013). *Steve jobs* (1st Simon & Schuster pbk. ed.). Simon & Schuster.

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

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology, 22*(3), 177–189. https://doi.org/10.1016/0022-1031(86)90022-3

Leung, S.-O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of Social Service Research, 37*(4), 412–421. https://doi.org/10.1080/01488376.2011.580697

Lord, C., Ross, L., & Lepper, M. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*, 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

Maglio, S. J., & Reich, T. (2019). Feeling certain: Gut choice, the true self, and attitude certainty. *Emotion, 19*(5), 876–888. https://doi.org/10.1037/emo0000490

Marin, P. M., Lindeman, M., & Svedholm-Häkkinen, A. M. (2024). Susceptibility to poor arguments: The interplay of cognitive sophistication and attitudes. *Memory & Cognition, 52*(7), 1579–1596. https://doi.org/10.3758/s13421-024-01564-1

Mercier, H. (2021). How good are we at evaluating communicated information? *Royal Institute of Philosophy Supplement, 89*, 257–272. https://doi.org/10.1017/S1358246121000096

Mercier, H. (2022). Confirmation bias – Myside bias. In *Cognitive illusions* (3rd ed.). Routledge.

Mercier, H., & Sperber, D. (2009). Intuitive and reflective inferences. In J. Evans, & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (1st ed., pp. 149–170). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0007.

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(2), 57–74. https://doi.org/10.1017/S0140525X10000968

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press. https://www.jstor.org/stable/j.ctv2sp3dd8.

Meyer, A., & Frederick, S. (2022). *Forming and revising intuitions* (SSRN Scholarly Paper No. 4039414). https://doi.org/10.2139/ssrn.4039414

Mikels, J. A., Maglio, S. J., Reed, A. E., & Kaplowitz, L. J. (2011). Should I go with my gut? Investigating the benefits of emotion-focused decision making. *Emotion, 11*(4), 743–753. https://doi.org/10.1037/a0023986

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

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

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*, 1154–1170. https://doi.org/10.1037/xlm0000372

Oktar, K., & Lombrozo, T. (2022). Deciding to be authentic: Intuition is favored over deliberation when authenticity matters. *Cognition, 223*, Article 105021. https://doi.org/10.1016/j.cognition.2022.105021

Pennycook, G. (2022). *A framework for understanding reasoning errors: From fake news to climate change and beyond*. PsyArXiv. https://doi.org/10.31234/osf.io/j3w7d.

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

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, The Cognitive Science of Political Thought, 188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Raoelison, M., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. *Judgment and Decision making, 14*(2), 170–178. https://doi.org/10.1017/S1930297500003405

Raoelison, M., Keime, M., & De Neys, W. (2021). Think slow, then fast: Does repeated deliberation boost correct intuitive responding? *Memory & Cognition, 49*(5), 873–883. https://doi.org/10.3758/s13421-021-01140-x

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science, 22*(4), 259–264. https://doi.org/10.1177/0963721413480174

Stieger, S., & Reips, U.-D. (2016). A limitation of the cognitive reflection test: Familiarity. *PeerJ, 4*, Article e2395. https://doi.org/10.7717/peerj.2395

Sultan, M., Tump, A. N., Ehmann, N., Lorenz-Spreen, P., Hertwig, R., Gollwitzer, A., & Kurvers, R. H. J. M. (2024). Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proceedings of the National Academy of Sciences of the United States of America, 121*(47), Article e2409329121. https://doi.org/10.1073/pnas.2409329121

Svedholm-Häkkinen, A. M., Forzani, E., Coiro, J., & Kiili, C. (2025). Online credibility evaluation skills in upper secondary students: The role of grade level, argument evaluation, and analytic thinking dispositions. *Learning and Individual Differences, 118*, Article 102640. https://doi.org/10.1016/j.lindif.2025.102640

Svedholm-Häkkinen, A. M., & Hietanen, M. (2024). On being drawn to different types of arguments: A mouse-tracking study. *Thinking & Reasoning*, 1–26. https://doi.org/10.1080/13546783.2024.2370074

Svedholm-Häkkinen, A. M., Hietanen, M., & Baron, J. (2024). Individual differences in argument strength discrimination. *Argumentation, 38*(2), 141–167. https://doi.org/10.1007/s10503-023-09620-x

Svedholm-Häkkinen, A. M., & Kiikeri, M. (2022). Cognitive miserliness in argument literacy? Effects of intuitive and analytic thinking on recognizing fallacies. *Judgment and Decision making, 17*(2), 331–361. https://doi.org/10.1017/S193029750000913X

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50*(3), 755–769.

Tanujaya, B., Prahmana, R. C. I., & Mumu, J. (2022). Likert scale in social sciences research: Problems and difficulties. *FWU Journal of Social Sciences*, 89–101. https://doi.org/10.51709/19951272/Winter2022/7

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

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

Walco, D. K., & Risen, J. L. (2017). The empirical case for acquiescing to intuition. *Psychological Science, 28*(12), 1807–1820. https://doi.org/10.1177/0956797617723377

Wolfe, M. B., & Kurby, C. A. (2017). Belief in the claim of an argument increases perceived argument soundness. *Discourse Processes, 54*(8), 599–617. https://doi.org/10.1080/0163853X.2015.1137446

Wu, H., & Leung, S.-O. (2017). Can Likert scales be treated as interval scales?—A simulation study. *Journal of Social Service Research, 43*(4), 527–532. https://doi.org/10.1080/01488376.2017.1329775

Supplementary Material

A. Items

Problems:

Problem type	Item number	Problem statement	Response option 1	Response option 2	Response option 3	Response option 4
C	1	In a building residents have 370 dogs and cats in total. There are 300 more dogs than cats. How many cats are there?	7	35	70	105
C	2	A store manager has bought 310 bananas and kiwis in total. There are 300 more bananas than kiwis. How many kiwis are there?	1	5	10	15
C	3	A store is advertising 220 coffee makers and toasters. There are 200 more coffee makers than toasters. How many toasters are there?	5	10	20	30
NC	4	A music school is renting 170 guitars and harps in total. There are 100 guitars. How many harps are there in this school?	7	35	70	105
NC	5	A national park has 650 roses and lotus flowers in total. There are 600 roses. How many lotus flowers are there in this park?	5	25	50	75
NC	6	A music store has 210 saxophones and flutes in total. There are 200 saxophones. How many flutes are there?	1	5	10	15

Arguments:

Problem type	Item number	Argument label	Argument response	Argument justification	Argument	Word count	Number of characters
C	1	A	Correct	Implicit	There are 35 cats, because the response that readily comes to mind when considering the problem is 35.	18	85
C	1	B	Correct	Explicit valid	There are 35 cats, because if there are 300 more dogs than cats among 370 pets, there must be 335 dogs and 35 cats.	24	92
C	1	C	Correct	Explicit invalid	There are 35 cats, because 100 dogs minus 65 is 35, so the solution must be that there are 35 cats in total.	23	86
C	1	D	Incorrect heuristic	Implicit	There are 70 cats, because the response that readily comes to mind when considering the problem is 70.	18	85

C	1	E	Incorrect heuristic	Explicit valid	There are 70 cats, because subtracting 300 from the total of 370 dogs and cats amounts to a total of 70 cats.	22	88
C	1	F	Incorrect heuristic	Explicit invalid	There are 70 cats, because if there are 300 dogs, then minus 200 is 100, and 100 dogs minus 30 amounts to 70 cats.	24	91
C	1	G	Incorrect random	Explicit invalid	There are 105 cats, because the 300 more dogs minus 230 is 70, and adding 35 amounts to a total of 105 cats.	23	86
C	2	A	Correct	Implicit	There are 5 kiwis, because the response that immediately comes to mind when considering the problem is 5.	18	88
C	2	B	Correct	Explicit valid	There are 5 kiwis, because if there are 300 more bananas than kiwis among 310 fruits, there must be 305 bananas and 5 kiwis.	24	101
C	2	C	Correct	Explicit invalid	There are 5 kiwis, because 200 bananas minus 195 is 5, so the solution must be that there are 5 kiwis in total.	23	89
C	2	D	Incorrect heuristic	Implicit	There are 10 kiwis, because the response that immediately comes to mind when considering the problem is 10.	18	90
C	2	E	Incorrect heuristic	Explicit valid	There are 10 kiwis, because subtracting 300 from the total of 310 bananas and kiwis amounts to a total of 10 kiwis.	22	94
C	2	F	Incorrect heuristic	Explicit invalid	There are 10 kiwis, because if there are 300 bananas, then minus 200 is 100, and 100 bananas minus 90 amounts to 10 kiwis.	24	99
C	2	G	Incorrect random	Explicit invalid	There are 15 kiwis, because the 300 more bananas minus 290 is 10, and adding 5 amounts to a total of 15 kiwis.	23	88
C	3	A	Correct	Implicit	There are 10 toasters, because the response that spontaneously comes to mind when considering the information in the problem is 10.	21	111
C	3	B	Correct	Explicit valid	There are 10 toasters, because if there are 200 more coffee makers than toasters among 220 items, there must be 210 coffee makers and 10 toasters.	26	121
C	3	C	Correct	Explicit invalid	There are 10 toasters, because 200 coffee makers minus 190 is 10, so the solution must be that there are 10 toasters in total.	24	103
C	3	D	Incorrect heuristic	Implicit	There are 20 toasters, because the response that spontaneously comes to mind when considering the information in the problem is 20.	21	111
C	3	E	Incorrect heuristic	Explicit valid	There are 20 toasters, because subtracting 200 from the total of	23	108

					220 coffee makers and toasters amounts to a total of 20 toasters.		
C	3	F	Incorrect heuristic	Explicit invalid	There are 20 toasters, because if there are 200 coffee makers, then minus 100 is 100, and 100 coffee makers minus 80 amounts to 20 toasters.	26	115
C	3	G	Incorrect random	Explicit invalid	There are 30 toasters, because the 200 more coffee makers minus 180 is 20, and adding 10 amounts to a total of 30 toasters.	24	100
NC	4	A	Correct	Implicit	There are 70 harps, because the response that readily comes to mind when considering the information in the problem is 70.	21	102
NC	4	B	Correct	Explicit valid	There are 70 harps, because subtracting 100 from the total of 170 guitars and harps amounts to a total of 70 harps.	22	94
NC	4	C	Correct	Explicit invalid	There are 70 harps, because if there are 200 guitars, then minus 100 is 100, and 100 guitars minus 30 amounts to 70 harps.	24	99
NC	4	D	Incorrect	Implicit	There are 35 harps, because the response that readily comes to mind when considering the information in the problem is 35.	21	102
NC	4	E	Incorrect	Explicit valid	There are 35 harps, because if there are 100 more guitars than harps among 170 items, there must be 135 guitars and 35 harps.	24	102
NC	4	F	Incorrect	Explicit invalid	There are 35 harps, because 100 guitars minus 65 is 35, so the solution must be that there are 35 harps in total.	23	91
NC	4	G	Incorrect random	Explicit invalid	There are 105 harps, because the 100 more guitars minus 30 is 70, and adding 35 amounts to a total of 105 harps.	23	90
NC	5	A	Correct	Implicit	There are 50 lotus flowers, because the response that immediately comes to mind when considering the information in the problem is 50.	22	113
NC	5	B	Correct	Explicit valid	There are 50 lotus flowers, because subtracting 600 from the total of 650 roses and lotus flowers amounts to a total of 50 lotus flowers.	25	113
NC	5	C	Correct	Explicit invalid	There are 50 lotus flowers, because if there are 600 roses, then minus 500 is 100, and 100 roses minus 50 amounts to 50 lotus flowers.	26	109
NC	5	D	Incorrect	Implicit	There are 25 lotus flowers, because the response that immediately comes to mind when considering the information in the problem is 25.	22	113
NC	5	E	Incorrect	Explicit valid	There are 25 lotus flowers, because if there are 600 more roses than lotus flowers among	27	121

					650 flowers, there must be 625 roses and 25 lotus flowers.		
NC	5	F	Incorrect	Explicit invalid	There are 25 lotus flowers, because 200 roses minus 175 is 25, so the solution must be that there are 25 lotus flowers in total.	25	104
NC	5	G	Incorrect random	Explicit invalid	There are 75 lotus flowers, because the 600 more roses minus 550 is 50, and adding 25 amounts to a total of 75 lotus flowers.	25	101
NC	6	A	Correct	Implicit	There are 10 flutes, because the response that spontaneously comes to mind when considering the problem is 10.	18	93
NC	6	B	Correct	Explicit valid	There are 10 flutes, because subtracting 200 from the total of 210 saxophones and flutes amounts to a total of 10 flutes.	22	100
NC	6	C	Correct	Explicit invalid	There are 10 flutes, because if there are 200 saxophones, then minus 100 is 100, and 100 saxophones minus 90 amounts to 10 flutes.	24	107
NC	6	D	Incorrect	Implicit	There are 5 flutes, because the response that spontaneously comes to mind when considering the problem is 5.	18	91
NC	6	E	Incorrect	Explicit valid	There are 5 flutes, because if there are 200 more saxophones than flutes among 210 items, there must be 205 saxophones and 5 flutes.	24	109
NC	6	F	Incorrect	Explicit invalid	There are 5 flutes, because 200 saxophones minus 195 is 5, so the solution must be that there are 5 flutes in total.	23	94
NC	6	G	Incorrect random	Explicit invalid	There are 15 flutes, because the 200 more saxophones minus 190 is 10, and adding 5 amounts to a total of 15 flutes.	23	93

Example of fast-intuitive argument evaluation trials:

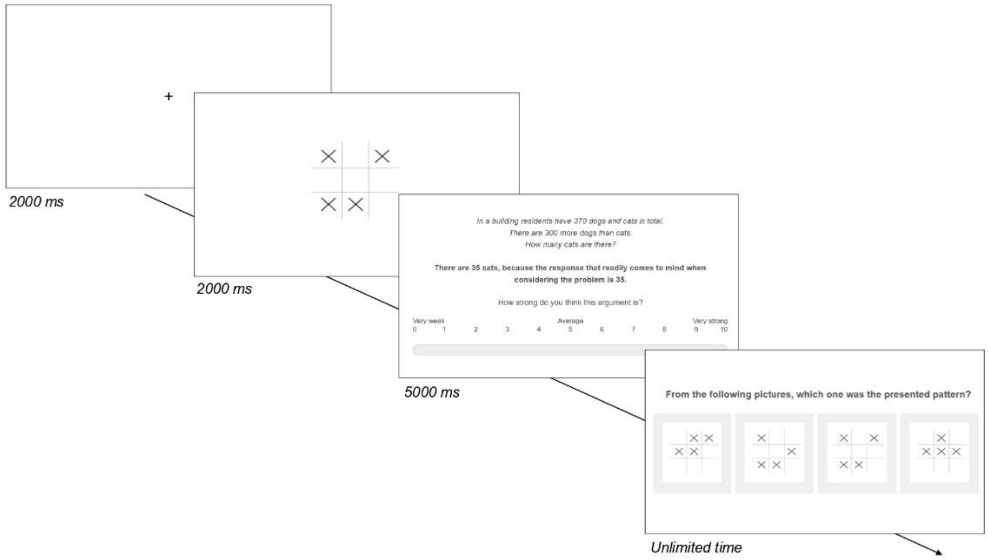

Fig.S1. Time course of a fast-intuitive evaluation trial in Study 2.

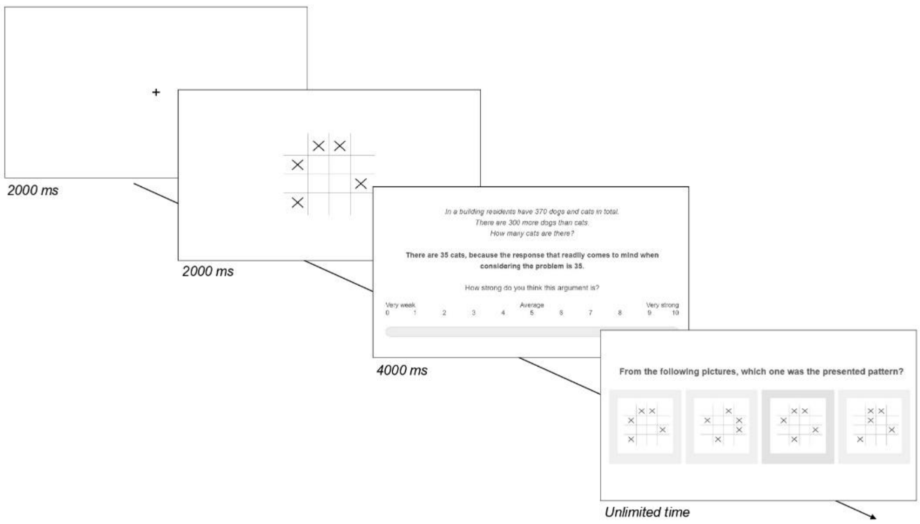

Fig.S2. Time course of a fast-intuitive evaluation trial in Study 3.

B. Reasoning problems

Table S1. Mean Accuracy by Condition for each Study (Conflict problems)

Study	Condition	Mean accuracy	SE
Study 1	Slow	32.68	4.31
Study 2	Fast	16.00	2.87
Study 2	Slow	36.34	3.92
Study 3	Fast	15.45	3.04
Study 3	Slow	32.61	3.72

Table S2. Mean Accuracy by Condition for each Study (No-conflict problems)

Study	Condition	Mean accuracy	SE
Study 1	Slow	99.35	0.46
Study 2	Fast	99.20	0.46
Study 2	Slow	98.50	0.93
Study 3	Fast	98.10	0.97
Study 3	Slow	98.55	0.58

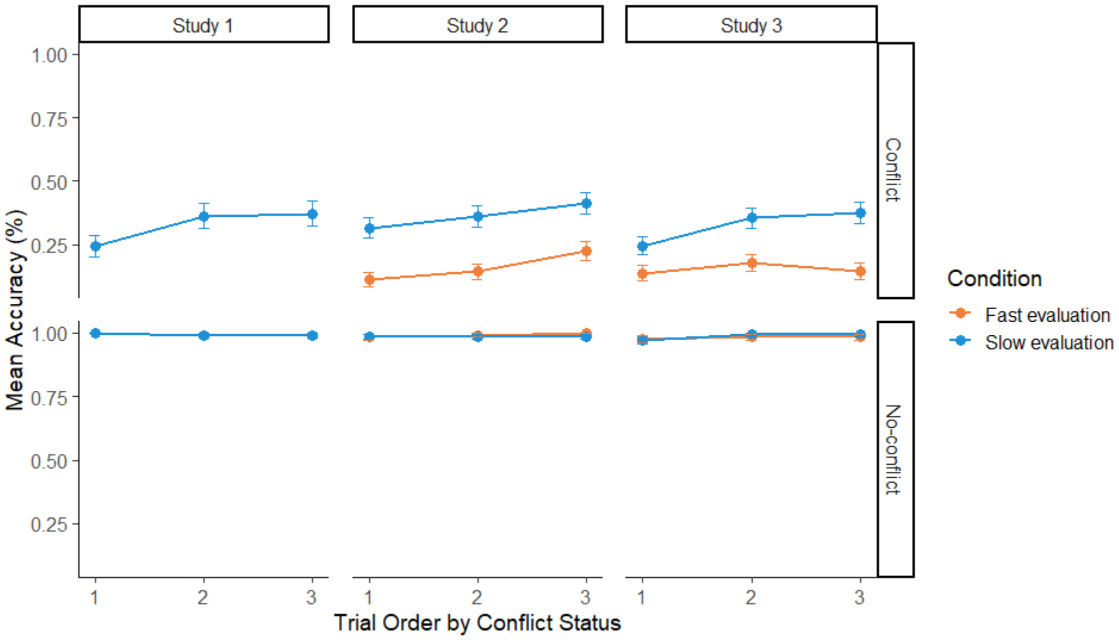

Fig.S3. Mean accuracy on problem solving by trial order, condition and problem status in each study. Error bars are standards errors of the mean.

If participants only answer one of the conflict problems correctly, it could suggest that they have guessed the correct answer. Thus, we computed the probability that a participant gets two conflict problems correct and three conflict problems correct given they got one conflict problem correct to see how many participants potentially guessed the correct answer vs used the correct reasoning strategy. However, after running the suggested analyses, the probability that a participant gets two or three conflict problems correct given they got one was high across studies and conditions (see Table S3).

Table S3. Summary of participant accuracy over the different conflict problems

Study	Condition	% with 2 or 3 correct (given >=1)	% with exactly 2 correct (given >=1)	% with exactly 3 correct (given >=1)
Study 1	Slow	94.87	33.33	61.54
Study 2	Fast	60.00	20.00	40.00
Study 2	Slow	89.29	19.64	69.64
Study 3	Fast	91.30	34.78	56.52
Study 3	Slow	90.57	26.42	64.15

Note that the inclusion of no-conflict problems, and the contrast between accuracy on conflict and on no-conflict problems, provides a first indication that people are not guessing overall. If they were, we would expect similar accuracy rates in the two types of problems, which is not the case.

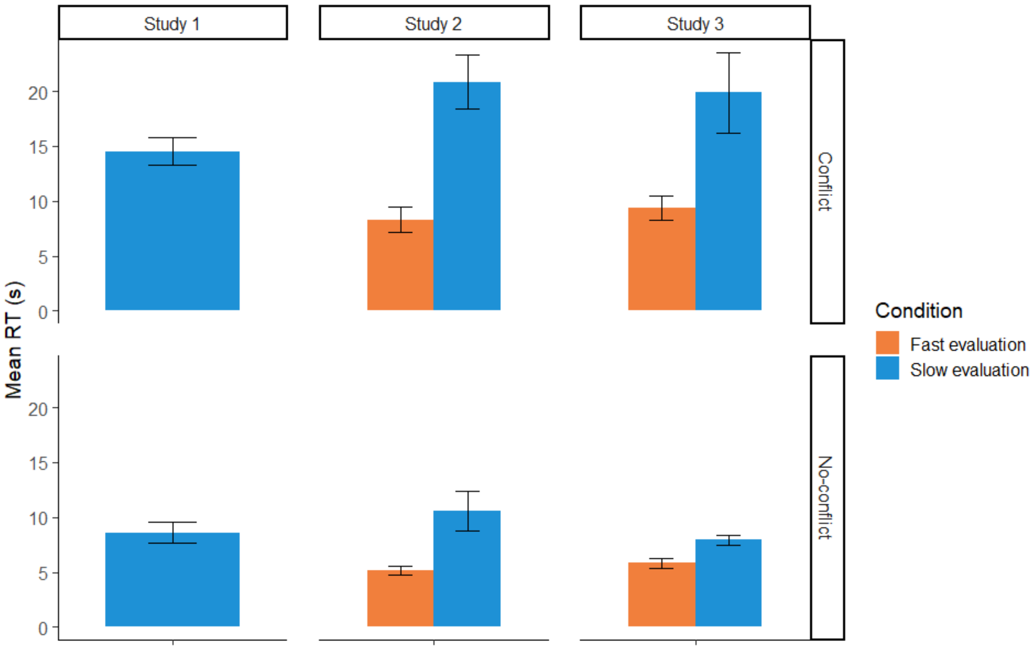

Fig.S4. Mean response time on problem solving by condition and problem status in each study. Error bars are standards errors of the mean.

On conflict problems, participants in the slow deliberate condition spent significantly more time to solve the problem than participants in the fast intuitive condition, in both Study 2, $F(1) = 197.9$, $p < .001$, and Study 3, $F(1) = 59.52$, $p < .001$.

As pointed out by a reviewer, given participants largely solved the conflict problems incorrectly, one might wonder whether these participants who got the problem wrong were deliberating. We thus compared the RTs on reasoning problems and on argument evaluations of participants who correctly solved the problems and of those who did not. Concerning the response time for problem solving, analyses show that those who correctly responded to the problems indeed took longer on average to solve a problem than those who gave incorrect answers (see Table S4 below). However, note that participants in the slow-deliberate condition who got problems wrong still took significantly longer to solve them than participants in the fast-intuitive condition who also got problems wrong (Study 2: $t(203) = 3.17$, $p = .002$; Study 3: $t(143) = 3.74$, $p < .001$), suggesting that they might still have engaged in (more) deliberation (although in any case, participants were not under constraint when solving the problems). Concerning response times for argument evaluations, analyses showed that participants in the slow deliberative condition who got the answer correct did not have significantly different response times than participants who got it wrong (Study 1: $t(113) = 1.79$, $p = .076$; Study 2: $t(148) = 1.83$, $p = .069$; Study 3: $t(153) = 1.88$, $p = .062$). This suggests that even those who got the problem wrong were taking their time to evaluate the arguments, or at least about the same time as those who got the problem right, which is our key concern.

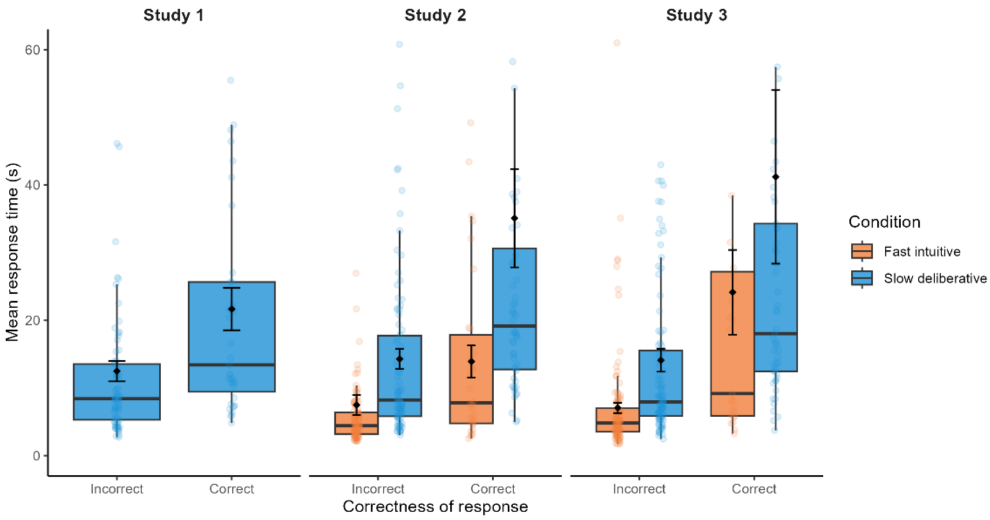

Fig.S5. Boxplots of response times on reasoning problems by response correctness and condition, for conflict problems. Diamonds show the mean for each subgroup, error bars are standards errors of the mean.

Table S4. T-tests for RT on correctly vs incorrectly solved conflict problems (slow condition)

Study	t_statistic	df	p_value
Study 1	-2.644	56.244	0.011
Study 2	-2.806	59.632	0.007
Study 3	-2.093	53.828	0.041

Table S5. T-tests for RT on correctly vs incorrectly solved conflict problems (fast condition)

Study	t_statistic	df	p_value
Study 2	-2.297	53.557	0.026
Study 3	-2.712	22.654	0.012

C. Argument ratings

Conflict problems

Table S6. Summary of number of participants and trials for each argument type in conflict problems

Study	Condition	Participant Response	Argument	Number of Participants	Number of Trials
Study 1	Slow	Incorrect	Correct implicit	78	206
			Correct explicit valid	78	206
			Correct explicit invalid	78	206
			Incorrect (heuristic) implicit	78	206
			Incorrect (heuristic) explicit valid	78	206
			Incorrect (heuristic) explicit invalid	78	206
			Incorrect control invalid	78	206
		Correct	Correct implicit	39	100
			Correct explicit valid	39	100
			Correct explicit invalid	39	100
			Incorrect (heuristic) implicit	39	100
			Incorrect (heuristic) explicit valid	39	100
			Incorrect (heuristic) explicit invalid	39	100
			Incorrect control invalid	39	100
Study 2	Fast	Incorrect	Correct implicit	113	275
			Correct explicit valid	107	235
			Correct explicit invalid	108	219
			Incorrect (heuristic) implicit	112	257
			Incorrect (heuristic) explicit valid	111	268
			Incorrect (heuristic) explicit invalid	109	264
			Incorrect control invalid	110	269
		Correct	Correct implicit	29	54
			Correct explicit valid	28	46
			Correct explicit invalid	28	48
			Incorrect (heuristic) implicit	29	54
			Incorrect (heuristic) explicit valid	30	58
			Incorrect (heuristic) explicit invalid	29	55
			Incorrect control invalid	30	50
	Slow	Incorrect	Correct implicit	94	254
			Correct explicit valid	94	254
			Correct explicit invalid	94	254
			Incorrect (heuristic) implicit	94	254
			Incorrect (heuristic) explicit valid	94	254
			Incorrect (heuristic) explicit invalid	94	254
			Incorrect control invalid	94	254
		Correct	Correct implicit	56	145
			Correct explicit valid	56	145
			Correct explicit invalid	56	145
			Incorrect (heuristic) implicit	56	145

Table S6. Summary of number of participants and trials for each argument type in conflict problems

Study	Condition	Participant Response	Argument	Number of Participants	Number of Trials
			Incorrect (heuristic) explicit valid	56	145
			Incorrect (heuristic) explicit invalid	56	145
			Incorrect control invalid	56	145
Study 3	Fast	Incorrect	Correct implicit	103	201
			Correct explicit valid	101	185
			Correct explicit invalid	103	201
			Incorrect (heuristic) implicit	97	187
			Incorrect (heuristic) explicit valid	99	193
			Incorrect (heuristic) explicit invalid	102	197
			Incorrect control invalid	102	228
		Correct	Correct implicit	20	35
			Correct explicit valid	18	32
			Correct explicit invalid	23	45
			Incorrect (heuristic) implicit	22	42
			Incorrect (heuristic) explicit valid	23	45
			Incorrect (heuristic) explicit invalid	21	42
			Incorrect control invalid	22	49
	Slow	Incorrect	Correct implicit	104	279
			Correct explicit valid	104	279
			Correct explicit invalid	104	279
			Incorrect (heuristic) implicit	104	279
			Incorrect (heuristic) explicit valid	104	279
			Incorrect (heuristic) explicit invalid	104	279
			Incorrect control invalid	104	279
		Correct	Correct implicit	53	135
			Correct explicit valid	53	135
			Correct explicit invalid	53	135
			Incorrect (heuristic) implicit	53	135
			Incorrect (heuristic) explicit valid	53	135
			Incorrect (heuristic) explicit invalid	53	135
			Incorrect control invalid	53	135

For ease of reading, arguments were labelled as follows:

A = Implicit correct

B = Explicit valid correct

C = Explicit invalid correct

D = Implicit incorrect (heuristic)

E = Explicit valid incorrect (heuristic)

F = Explicit invalid incorrect (heuristic)

G = Control invalid incorrect (high foil)

Study 1:

Correct responders:

Table S7. ANOVA on correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.993	75.736	4.277	141.798	0.698	<.001

Table S8. Post-hoc pairwise comparisons for correct response arguments among correct responders

contrast	estimate	SE	df	t.ratio	d	p.value
A - B	-6.615	0.454	38	-14.570	2.333	<.001
A - C	0.389	0.477	38	0.815	.130	0.696
B - C	7.004	0.471	38	14.874	2.382	<.001

Table S9. ANOVA on incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.878	71.372	1.999	9.263	0.063	<.001

Table S10. Post-hoc pairwise comparisons for incorrect response arguments among correct responders

contrast	estimate	SE	df	t.ratio	d	p.value
D - E	-1.111	0.287	38	-3.872	.620	0.002
D - F	-0.222	0.169	38	-1.312	.210	0.561
D - G	0.073	0.241	38	0.301	.048	0.990
E - F	0.889	0.282	38	3.153	.505	0.016
E - G	1.184	0.343	38	3.449	.552	0.007
F - G	0.295	0.137	38	2.155	.345	0.155

Incorrect responders:

Table S11. ANOVA on correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.429	109.997	4.83	23.227	0.122	<.001

Table S12. Post-hoc pairwise comparisons for correct response arguments among incorrect responders

contrast	estimate	SE	df	t.ratio	d	p.value
A - B	-1.912	0.358	77	-5.343	.605	<.001
A - C	-0.374	0.186	77	-2.015	.228	0.116
B - C	1.538	0.321	77	4.798	.543	<.001

Table S13. ANOVA for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.743	211.239	6.492	126.847	0.479	<.001

Table S14. Post-hoc pairwise comparisons for incorrect response arguments among incorrect responders

contrast	estimate	SE	df	t.ratio	d	p.value
D - E	-3.868	0.419	77	-9.240	1.046	<.001
D - F	1.152	0.391	77	2.943	.333	0.022
D - G	3.588	0.439	77	8.178	.926	<.001
E - F	5.019	0.398	77	12.615	1.428	<.001
E - G	7.455	0.358	77	20.827	2.358	<.001
F - G	2.436	0.326	77	7.479	.847	<.001

Study 2:

Correct responders:

Table S15. ANOVA on Fast responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.972	47.332	6.347	14.702	0.234	<.001

Table S16. ANOVA on Slow responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.946	107.03	3.127	264.296	0.688	<.001

Table S17. Post-hoc pairwise comparisons for correct response arguments among correct responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
A - B	Fast	-3.020	0.605	79	-4.993	-.850	<.001
A - C	Fast	0.540	0.558	79	0.968	.044	0.599
B - C	Fast	3.560	0.537	79	6.635	.900	<.001
A - B	Slow	-6.149	0.404	79	-15.214	-2.316	<.001
A - C	Slow	0.762	0.373	79	2.045	.314	0.108
B - C	Slow	6.911	0.358	79	19.277	2.992	<.001

Table S18. ANOVA on Fast responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.662	71.869	1.844	1.61	0.017	0.199

Table S19. ANOVA on Slow responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.345	128.956	1.989	16.785	0.1	<.001

Table S20. Post-hoc pairwise comparisons on incorrect response arguments among correct responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
D - E	Fast	-0.708	0.322	82	-2.198	-.420	0.133
D - F	Fast	-0.161	0.288	82	-0.558	-.099	0.944
D - G	Fast	-0.196	0.376	82	-0.522	-.076	0.954
E - F	Fast	0.548	0.303	82	1.810	.318	0.276
E - G	Fast	0.512	0.405	82	1.263	.234	0.589
F - G	Fast	-0.036	0.306	82	-0.117	-.017	0.999
D - E	Slow	-1.152	0.228	82	-5.053	-.647	<.001
D - F	Slow	-0.116	0.204	82	-0.570	-.079	0.941
D - G	Slow	0.467	0.266	82	1.756	.238	0.302
E - F	Slow	1.036	0.214	82	4.841	.638	<.001
E - G	Slow	1.619	0.287	82	5.647	.724	<.001
F - G	Slow	0.583	0.216	82	2.699	.432	0.041

Table S21. ANOVA: Correct vs Incorrect response arguments - Fast evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	29	3.158	75.361	0.554	<.001

Table S22. ANOVA: Correct vs Incorrect response arguments - Slow evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	55	1.238	236.685	0.547	<.001

Incorrect responders:

Table S23. ANOVA on Fast responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.913	197.018	1.511	2.558	0.004	0.083

Table S24. ANOVA on Slow responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.655	153.9	3.011	16.965	0.058	<.001

Table S25. Post-hoc pairwise comparisons for correct response arguments among incorrect responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
A - B	Fast	-0.377	0.217	196	-1.736	-.221	0.252
A - C	Fast	-0.204	0.163	196	-1.252	-.121	0.636
B - C	Fast	0.173	0.196	196	0.882	.111	1.000
A - B	Slow	-1.216	0.228	196	-5.330	-.479	<.001
A - C	Slow	-0.119	0.171	196	-0.695	-.072	1.000
B - C	Slow	1.098	0.206	196	5.316	.458	<.001

Table S26. ANOVA on Fast responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.626	273.115	5.717	96.432	0.352	<.001

Table S27. ANOVA on Slow responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.69	250.135	6.448	170.967	0.514	<.001

Table S28. Post-hoc pairwise comparisons for incorrect response arguments among incorrect responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
D - E	Fast	-1.413	0.309	197	-4.578	-0.469	<.001
D - F	Fast	0.673	0.311	197	2.166	0.237	0.189
D - G	Fast	3.673	0.380	197	9.658	0.938	<.001
E - F	Fast	2.086	0.295	197	7.079	0.714	<.001
E - G	Fast	5.086	0.315	197	16.124	1.421	<.001

contrast	condition	estimate	SE	df	t.ratio	d	p.value
F - G	Fast	3.000	0.302	197	9.941	0.887	<.001
D - E	Slow	-3.794	0.326	197	-11.633	-1.081	<.001
D - F	Slow	1.163	0.328	197	3.541	0.324	0.003
D - G	Slow	4.035	0.402	197	10.040	1.017	<.001
E - F	Slow	4.957	0.311	197	15.920	1.520	<.001
E - G	Slow	7.830	0.333	197	23.488	2.492	<.001
F - G	Slow	2.872	0.319	197	9.005	1.020	<.001

Table S29. ANOVA: Correct vs Incorrect response arguments - Fast evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	112	4.878	134.79	0.36	<.001

Table S30. ANOVA: Correct vs Incorrect response arguments - Slow evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	93	4.425	204.982	0.532	<.001

Study 3:

Correct responders:

Table S31. ANOVA on Fast responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.164	16.298	3.204	11.264	0.109	0.003

Table S32. ANOVA on Slow responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.804	93.806	3.931	209.439	0.668	<.001

Table S33. Post-hoc pairwise comparisons for correct response arguments among correct responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
A - B	Fast	-2.367	0.740	66	-3.197	-0.902	0.006
A - C	Fast	-1.189	0.531	66	-2.241	-0.632	0.085

contrast	condition	estimate	SE	df	t.ratio	d	p.value
B - C	Fast	1.178	0.668	66	1.763	0.274	0.247
A - B	Slow	-6.160	0.394	66	-15.644	-2.104	<.001
A - C	Slow	0.604	0.282	66	2.139	0.276	0.108
B - C	Slow	6.764	0.355	66	19.035	2.405	<.001

Table S34. ANOVA on Fast responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.718	51.647	0.859	9.226	0.087	<.001

Table S35. ANOVA on Slow responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.971	102.484	1.699	14.359	0.103	<.001

Table S36. Post-hoc pairwise comparisons for incorrect response arguments among correct responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
D - E	Fast	-1.025	0.378	71	-2.714	-0.696	0.050
D - F	Fast	-1.025	0.290	71	-3.530	-0.738	0.004
D - G	Fast	0.025	0.272	71	0.092	-0.082	1.000
E - F	Fast	0.000	0.378	71	0.000	0.141	1.000
E - G	Fast	1.050	0.365	71	2.878	0.548	0.032
F - G	Fast	1.050	0.196	71	5.358	0.768	<.001
D - E	Slow	-0.959	0.232	71	-4.134	-0.521	<.001
D - F	Slow	0.028	0.178	71	0.159	-0.022	1.000
D - G	Slow	0.296	0.167	71	1.771	0.230	0.486
E - F	Slow	0.987	0.232	71	4.249	0.546	<.001
E - G	Slow	1.255	0.224	71	5.598	0.710	<.001
F - G	Slow	0.267	0.120	71	2.221	0.442	0.177

Table S37. ANOVA: Correct vs Incorrect response arguments - Fast evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	22	2.942	26.318	0.304	<.001

Table S38. ANOVA: Correct vs Incorrect response arguments - Slow evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	52	1.074	237.848	0.582	<.001

Incorrect responders:

Table S39.ANOVA on Fast responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.979	178.142	2.168	2.343	0.005	0.100

Table S40. ANOVA on Slow responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.329	136.897	4.956	26.344	0.089	<.001

Table S41. Post-hoc pairwise comparisons for correct response arguments among incorrect responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
A - B	Fast	-0.353	0.272	193	-1.298	-0.150	0.587
A - C	Fast	0.092	0.188	193	0.488	0.026	1.000
B - C	Fast	0.445	0.269	193	1.653	0.219	0.300
A - B	Slow	-1.633	0.255	193	-6.412	-0.546	<.001
A - C	Slow	-0.107	0.176	193	-0.612	-0.078	1.000
B - C	Slow	1.526	0.252	193	6.057	0.511	<.001

Table S42. ANOVA on Fast responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.256	176	7.233	40.587	0.231	<.001

Table S43. ANOVA on Slow responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.608	268.64	6.582	162.989	0.471	<.001

Table S44. Post-hoc pairwise comparisons for incorrect response arguments among incorrect responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
D - E	Fast	-0.397	0.325	181	-1.221	-0.222	1.000
D - F	Fast	1.502	0.356	181	4.217	0.498	<.001
D - G	Fast	3.285	0.429	181	7.649	0.874	<.001
E - F	Fast	1.899	0.375	181	5.060	0.625	<.001
E - G	Fast	3.681	0.419	181	8.787	0.932	<.001
F - G	Fast	1.783	0.343	181	5.197	0.529	<.001
D - E	Slow	-3.176	0.283	181	-11.223	-0.936	<.001
D - F	Slow	1.846	0.310	181	5.947	0.547	<.001
D - G	Slow	3.877	0.374	181	10.357	1.022	<.001
E - F	Slow	5.022	0.327	181	15.357	1.408	<.001
E - G	Slow	7.053	0.365	181	19.315	2.009	<.001
F - G	Slow	2.030	0.299	181	6.791	0.806	<.001

Table S45. ANOVA: Correct vs Incorrect response arguments - Fast evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	108	5.314	90.893	0.294	<.001

Table S46. ANOVA: Correct vs Incorrect response arguments - Slow evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	103	5.467	138.203	0.455	<.001

Contrast of Fast and Slow ratings

Table S47: Post-hoc pairwise comparisons of fast vs slow ratings

Study	Participant response	Argument	Estimate (Fast-Slow)	SE	df	t.ratio	d	p.value
Study 2	Incorrect	Correct implicit	1.648	0.308	196	5.360	0.808	<.001
		Correct explicit valid	0.809	0.403	196	2.008	0.306	0.046
		Correct explicit invalid	1.733	0.305	196	5.678	0.845	<.001
		Incorrect (heuristic) implicit	1.343	0.462	197	2.909	0.392	0.004
		Incorrect (heuristic) explicit valid	-1.038	0.325	197	-3.192	-0.511	0.002
		Incorrect (heuristic) explicit invalid	1.834	0.401	197	4.571	0.629	<.001
		Incorrect control invalid	1.706	0.282	197	6.044	0.899	<.001
	Correct	Correct implicit	2.357	0.684	79	3.447	0.766	<.001
		Correct explicit valid	-0.772	0.434	79	-1.776	-0.489	0.080
		Correct explicit invalid	2.579	0.565	79	4.568	1.207	<.001
		Incorrect (heuristic) implicit	0.661	0.457	82	1.446	0.322	0.152
		Incorrect (heuristic) explicit valid	0.217	0.542	82	0.401	0.021	0.690
		Incorrect (heuristic) explicit invalid	0.705	0.369	82	1.914	0.424	0.059
		Incorrect control invalid	1.324	0.344	82	3.854	0.824	<.001
Study 3	Incorrect	Correct implicit	1.945	0.312	193	6.239	0.942	<.001
		Correct explicit valid	0.666	0.435	193	1.531	0.228	0.128
		Correct explicit invalid	1.746	0.319	193	5.482	0.810	<.001
		Incorrect (heuristic) implicit	1.511	0.460	181	3.287	0.501	0.001
		Incorrect (heuristic) explicit valid	-1.269	0.406	181	-3.125	-0.477	0.002
		Incorrect (heuristic) explicit invalid	1.855	0.427	181	4.346	0.687	<.001
		Incorrect control invalid	2.103	0.317	181	6.637	0.939	<.001
	Correct	Correct implicit	1.463	0.728	66	2.010	0.703	0.048
		Correct explicit valid	-2.331	0.660	66	-3.530	-1.098	<.001
		Correct explicit invalid	3.256	0.658	66	4.946	1.697	<.001
		Incorrect (heuristic) implicit	1.100	0.411	71	2.677	0.777	0.009
		Incorrect (heuristic) explicit valid	1.166	0.500	71	2.330	0.660	0.023
		Incorrect (heuristic) explicit invalid	2.153	0.353	71	6.108	1.460	<.001
		Incorrect control invalid	1.370	0.285	71	4.816	1.306	<.001

Response times

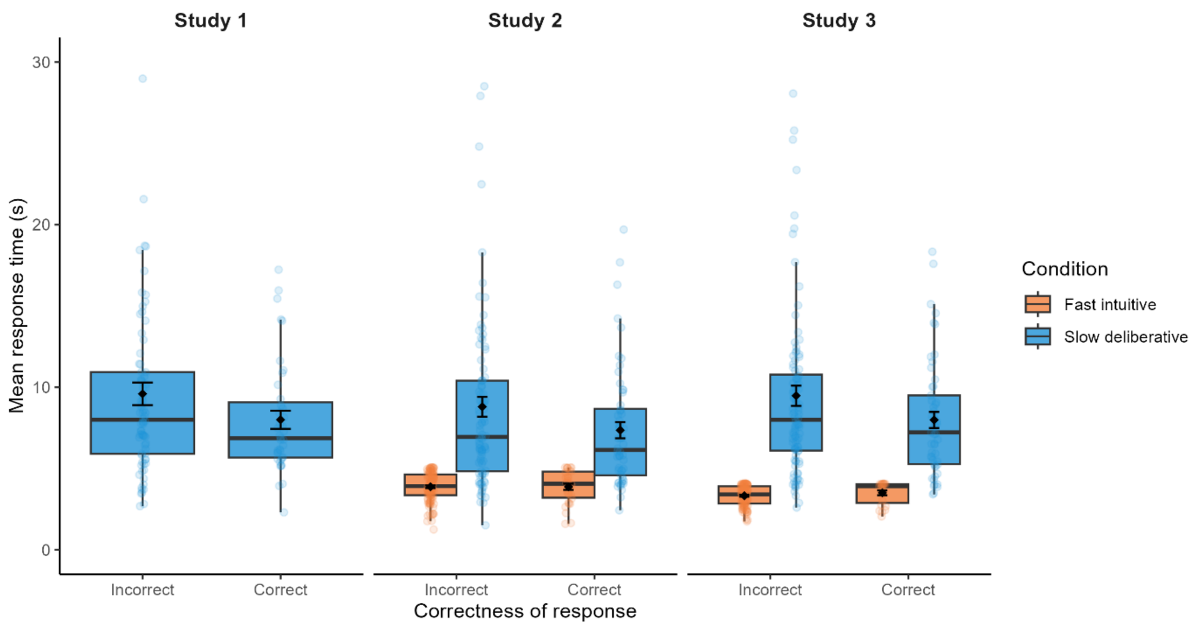

Fig.S6. Boxplots of response times on argument evaluations by response correctness and condition, for conflict problems. Diamonds show the mean for each subgroup, error bars are standards errors of the mean.

Table S48. T-tests for argument evaluation RT after correctly vs incorrectly solved reasoning problems (slow condition)

Study	t_statistic	df	p_value
Study 1	1.789	112.94	0.076
Study 2	1.8307	147.63	0.069
Study 3	1.8781	153.35	0.062

Table S49. T-tests for argument evaluation RT after correctly vs incorrectly solved reasoning problems (fast condition)

Study	t_statistic	df	p_value
Study 2	0.007	41.258	0.994
Study 3	-1.147	29.617	0.260

No-conflict problems

Study 1

The mean ratings by argument response and justification type on no-conflict problems are presented in Figure S5. Given accuracy rates were at ceiling, we did not analyze the few trials on which participants responded incorrectly. Starting with correct response arguments (left of the dashed line), participants gave the highest ratings to explicit valid arguments supporting the correct answer (M = 8.97, SEM = .13). Bonferroni-corrected comparison analyses showed that these were significantly higher than both implicit (M = 4.20, SEM = .22; $t(101) = 12.44$, $p < .001$, $d = 1.23$) and explicit invalid (M = 2.81, SEM = .17 ; $t(101) = 19.13$, $p < .001$, $d = 1.89$) correct arguments. Hence, in no-conflict problems reasoners also preferred stronger, explicit valid arguments. Overall participants rated correct response arguments higher than incorrect response arguments, as indicated by a repeated-measures ANOVA, $F(2.71, 273.31) = 258.05$, $p < .001$, $\eta_G^2 = .633$. All arguments supporting incorrect answers received very low ratings. Nevertheless, for completeness, note that even among these arguments, explicit ones received slightly higher ratings than implicit ($M_{explicit} = .97$, SEM = .09, $M_{implicit} = .66$, SEM = .08, $t(101) = 3.23$, $p = .01$, $d = .32$) as well as explicit invalid (M = .78, SEM = .08, $t(101) = 2.71$, $p = .047$, $d = .27$) and control invalid justifications (M = .72, SEM = .07, $t(101) = 3.68$, $p < .01$, $d = .36$).

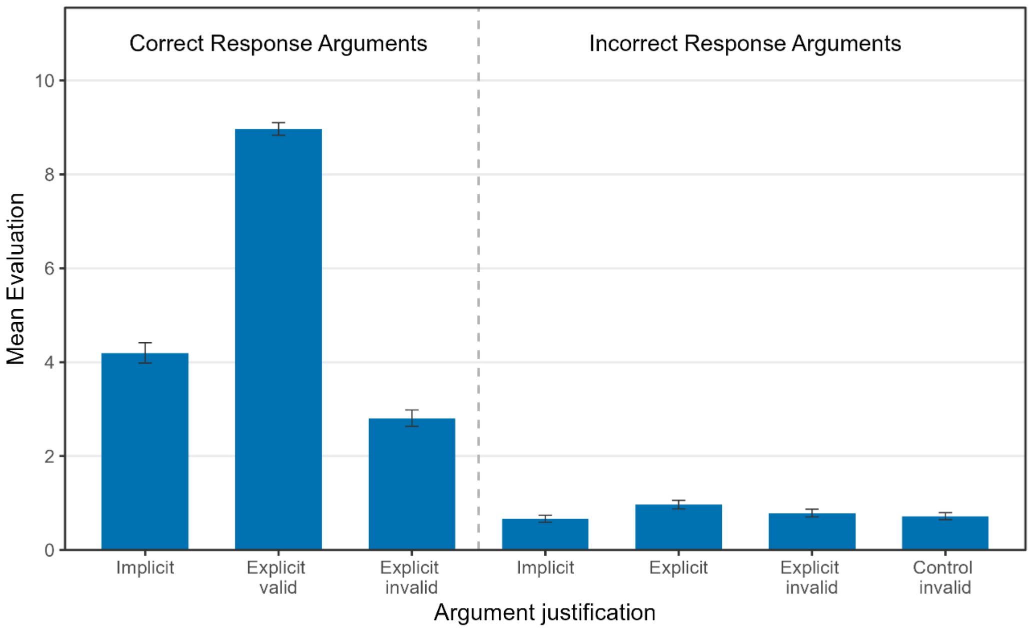

Fig.S7. Mean evaluations of arguments on no-conflict problems by argument response and justification (Study 1). Error bars are standard errors of the mean.

Table S50. Post-hoc pairwise comparisons for correct response arguments among correct responders (NC problems)

contrast	estimate	SE	df	t.ratio	d	p.value
A - B	-4.792	0.385	101	-12.442	-1.232	<.001
A - C	1.373	0.361	101	3.800	0.376	<.001
B - C	6.165	0.322	101	19.133	1.894	<.001

Table S51. Post-hoc pairwise comparisons for incorrect response arguments among correct responders (NC problems)

contrast	estimate	SE	df	t.ratio	d	p.value
D - E	-0.309	0.096	101	-3.234	-0.320	0.010
D - F	-0.123	0.084	101	-1.452	-0.144	0.898
D - G	-0.059	0.086	101	-0.685	-0.068	1.000
E - F	0.186	0.069	101	2.713	0.269	0.047
E - G	0.250	0.068	101	3.684	0.365	0.002
F - G	0.064	0.052	101	1.219	0.121	1.000

Studies 2 and 3

Mean ratings by argument response and justification type after no-conflict (control) problems are presented in Figure S6A (Study 2) and S6B (Study 3). Note that we report intuitive and deliberate rating results following correct responses, as there were almost no incorrect answers on these problems.

Starting with arguments supporting the correct response, a clear preference for explicit valid justifications was again observed in the slow deliberate condition of both studies. Participants gave the highest ratings to explicit valid arguments (Study 2: M = 9.28, SEM = .10; Study 3: M = 7.24, SEM = .18), significantly above all other correct arguments as confirmed by Bonferroni-corrected analyses (Study 2: all $17.1 < t < 23.3$, $p < .001$, $1.43 < d < 1.99$; Study 3: $17.21 < t < 25.13$, $p < .001$, $1.33 < d < 2.00$). In the fast intuitive condition, this pattern persisted but was attenuated. Repeated-measures ANOVAs on correct-response arguments revealed a significant interaction between condition and argument in both studies (Study 2: $F(1.98, 491.50) = 56.54$, $p < .001$, $\eta_G^2 = .095$; Study 3: $F(2.74, 611.57) = 37.03$, $p < .001$, $\eta_G^2 = .098$). Bonferroni-corrected contrasts showed that correct explicit valid arguments received lower ratings under intuitive than under deliberate evaluation (Study 2: $t(246) = -5.36$, $p < .001$, d = -

.70 ; Study 3: $t(223) = -7.14$, $p < .001$, $d = -.91$), while the remaining correct arguments received higher ratings under intuitive evaluation (Study 2: all $4.50 < t < 7.48$, $p < .01$, $.60 < d < .96$; Study 3: $3.01 < t < 7.44$, $p < .01$, $.46 < d < 1.12$). Importantly, the preference for explicit over implicit justifications remained significant under intuitive evaluation in both studies (Study 2: all $5.28 < t < 7.59$, $p < .001$, $.50 < d < .69$; Study 3: all $4.13 < t < 5.24$, $p < .001$, $.48 < d < .54$), though the distinctions were smaller than under deliberation.

Turning to the evaluation of incorrect-response arguments, patterns differed. In the slow deliberate condition, incorrect arguments with explicit valid justifications received slightly but significantly higher ratings than implicit or invalid ones (Study 2: all $2.98 < t < 4.23$, $p < .05$, $.34 < d < .43$; Study 3: all $2.31 < t < 4.67$, $p < .05$, $.23 < d < .40$). In contrast, under fast intuitive evaluation, these distinctions largely disappeared. All incorrect-response arguments received similar ratings, with no significant differences detected in Study 2: $F(2.74, 326.05) = 2.49$, $p = .066$, nor in Study 3: $F(2.84, 260.87) = .94$, $p = .42$. Thus, while the preference for explicit justifications remained under intuitive processing for correct-response arguments, it faded entirely for incorrect-response arguments. These findings further support the conclusion that sensitivity to justification strength diminishes when arguments are evaluated intuitively.

Note that in the slow deliberate condition, participants gave overall lower ratings to incorrect-response arguments than to correct-response arguments as confirmed by repeated-measures ANOVA, Study 2: $F(1, 131) = 651.86$, $p < .001$, $\eta_G^2 = .695$, Study 3: $F(1, 137) = 642.88$, $p < .001$, $\eta_G^2 = .664$, replicating the pattern from Study 1. This difference in ratings was slightly attenuated in the fast intuitive condition, as evidenced by significant interactions between condition and argument group, Study 2: $F(1, 255) = 5.55$, $p = .019$, $\eta_G^2 = .01$, Study 3: $F(1, 256) = 14.30$, $p < .001$, $\eta_G^2 = .027$. Yet, in the intuitive condition correct-response arguments still received overall higher ratings than incorrect-response arguments, Study 2: $F(1, 124) = 221.65$, $p < .001$, $\eta_G^2 = .464$, Study 3: $F(1, 119) = 125.10$, $p < .001$, $\eta_G^2 = .362$.

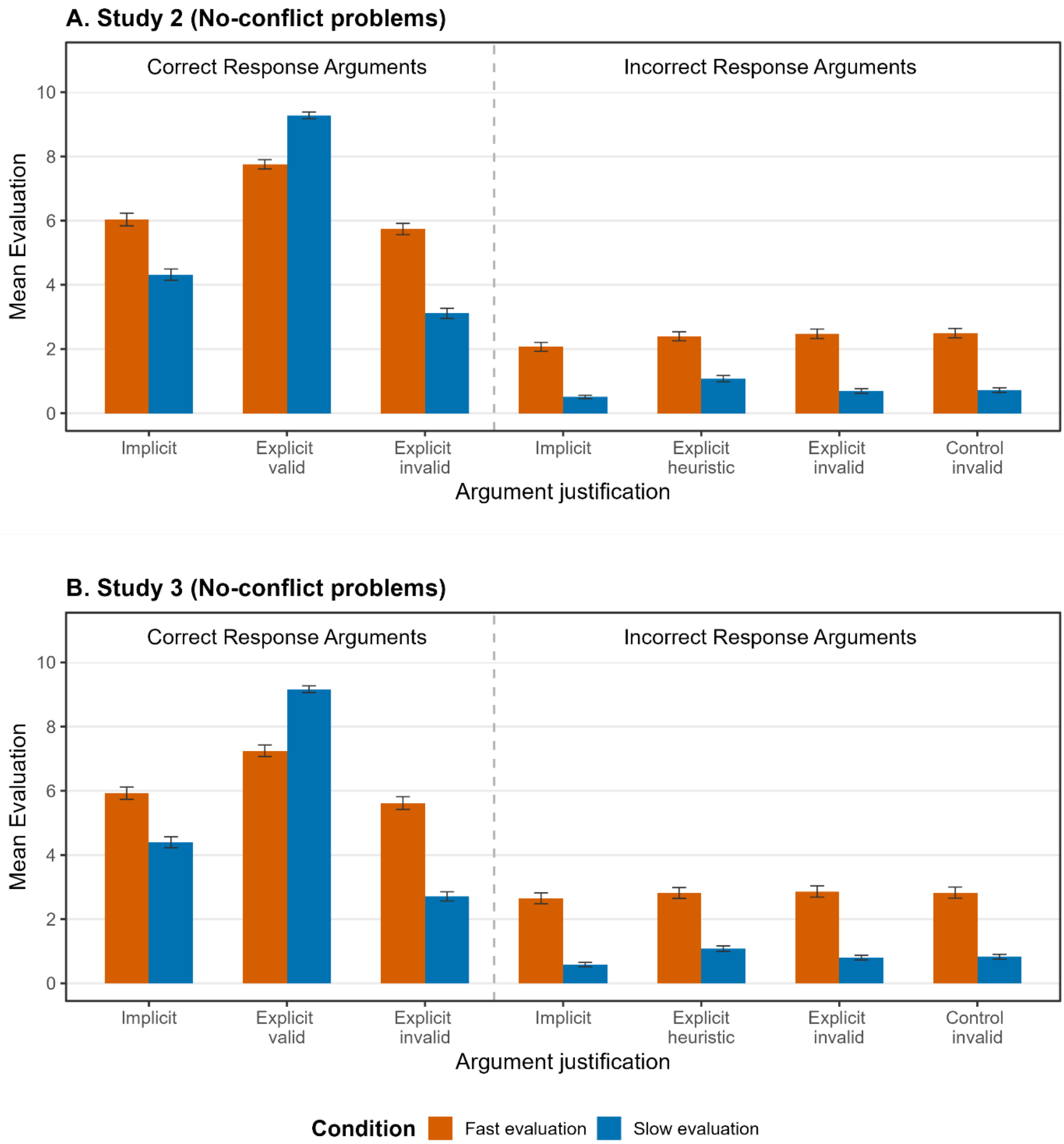

Fig.S8. Mean evaluation of arguments on no-conflict problems by condition in Study 2 (A) and Study 3 (B), separated by participant response to the problem and argument response correctness. Error bars are standard errors of the mean.

Table S52. Post-hoc pairwise comparisons for correct response arguments among correct responders (NC problems, Study 2)

contrast	condition	estimate	SE	df	t.ratio	d	p.value
A - B	Fast	-1.616	0.306	248	-5.281	- 0.504	<.001
A - C	Fast	0.521	0.297	248	1.753	0.168	0.243
B - C	Fast	2.137	0.282	248	7.586	0.685	<.001
A - B	Slow	-4.949	0.289	248	-17.108	-1.435	<.001
A - C	Slow	1.234	0.281	248	4.387	0.378	<.001
B - C	Slow	6.183	0.266	248	23.214	1.989	<.001

Table S53. Post-hoc pairwise comparisons on incorrect response arguments among correct responders (NC problems, Study 2)

contrast	condition	estimate	SE	df	t.ratio	d	p.value
D - E	Fast	-0.342	0.141	250	-2.430	-0.191	0.095
D - F	Fast	-0.212	0.123	250	-1.724	-0.128	0.516
D - G	Fast	-0.353	0.126	250	-2.809	- 0.205	0.032
E - F	Fast	0.129	0.107	250	1.212	0.068	1.000
E - G	Fast	-0.011	0.123	250	-0.090	- 0.011	1.000
F - G	Fast	-0.140	0.111	250	-1.265	- 0.062	1.000
D - E	Slow	-0.566	0.134	250	-4.220	-0.421	<.001
D - F	Slow	-0.183	0.118	250	-1.558	-0.184	0.723
D - G	Slow	-0.216	0.120	250	-1.803	-0.205	0.436
E - F	Slow	0.383	0.102	250	3.765	0.353	0.001
E - G	Slow	0.350	0.117	250	2.981	0.344	0.019
F - G	Slow	-0.033	0.106	250	-0.311	-0.066	1.000

Table S54. Post-hoc pairwise comparisons of Fast vs Slow ratings (NC problems, Study 2)

contrast	argument	estimate	SE	df	t.ratio	d	p.value
Fast - Slow	A	1.853	0.411	246	4.506	0.600	<.001
Fast - Slow	B	-1.465	0.273	246	-5.364	-0.702	<.001
Fast - Slow	C	2.598	0.347	246	7.479	0.951	<.001
Fast - Slow	D	1.534	0.204	246	7.528	0.962	<.001
Fast - Slow	E	1.375	0.247	246	5.576	0.742	<.001

contrast	argument	estimate	SE	df	t.ratio	d	p.value
Fast - Slow	F	1.602	0.222	246	7.205	0.951	<.001
Fast - Slow	G	1.730	0.241	246	7.187	0.942	<.001

Table S55. Post-hoc pairwise comparisons for correct response arguments among correct responders (NC problems, Study 3)

contrast	condition	estimate	SE	df	t.ratio	d	p.value
A - B	Fast	-1.324	0.320	239	-4.139	-0.488	<.001
A - C	Fast	0.231	0.289	239	0.801	0.073	1.000
B - C	Fast	1.555	0.297	239	5.234	0.536	<.001
A - B	Slow	-4.755	0.276	239	-17.212	-1.330	<.001
A - C	Slow	1.694	0.250	239	6.788	0.562	<.001
B - C	Slow	6.449	0.257	239	25.125	1.996	<.001

Table S56. Post-hoc pairwise comparisons on incorrect response arguments among correct responders (NC problems, Study 3)

contrast	condition	estimate	SE	df	t.ratio	d	p.value
D - E	Fast	-0.203	0.128	229	-1.581	-0.127	0.691
D - F	Fast	-0.093	0.105	229	-0.890	-0.104	1.000
D - G	Fast	-0.018	0.111	229	-0.162	-0.024	1.000
E - F	Fast	0.109	0.114	229	0.963	0.070	1.000
E - G	Fast	0.185	0.127	229	1.452	0.106	0.887
F - G	Fast	0.075	0.095	229	0.789	0.066	1.000
D - E	Slow	-0.490	0.105	229	-4.663	-0.395	<.001
D - F	Slow	-0.217	0.086	229	-2.528	-0.282	0.073
D - G	Slow	-0.249	0.091	229	-2.735	-0.298	0.040
E - F	Slow	0.273	0.093	229	2.929	0.257	0.022
E - G	Slow	0.242	0.104	229	2.315	0.230	0.129
F - G	Slow	-0.031	0.078	229	-0.401	-0.061	1.000

Table S57. Post-hoc pairwise comparisons of Fast vs Slow ratings (NC problems, Study 3)

contrast	argument	estimate	SE	df	t.ratio	d	p.value
Fast - Slow	A	1.299	0.432	223	3.010	0.465	0.003

contrast	argument	estimate	SE	df	t.ratio	d	p.value
Fast - Slow	B	-2.197	0.308	223	-7.141	-0.914	<.001
Fast - Slow	C	2.678	0.360	223	7.439	1.116	<.001
Fast - Slow	D	2.127	0.245	223	8.674	1.138	<.001
Fast - Slow	E	1.851	0.268	223	6.899	0.898	<.001
Fast - Slow	F	1.975	0.241	223	8.182	1.092	<.001
Fast - Slow	G	1.871	0.238	223	7.844	1.019	<.001

D. Control analyses

According to the preregistration, we performed separate data analyses with and without participants who were familiar with the classic version of the bat-and-ball problem. Both analyses yielded similar results. Hence, as planned in the preregistration we report in the manuscript the results of the full analysis including all participants. We summarise below where patterns slightly differed in the analyses excluding participants who were familiar with the classic bat-and-ball problem.

Unsurprisingly, when removing participants familiar with the classic bat-and-ball problem, response accuracy on conflict problems was reduced compared to the results including all participants. This was true for all studies, in both conditions. However, accuracy patterns remained similar across studies with those observed in the main analysis, with participants from the fast-intuitive conditions showing a lower accuracy than those from the slow-deliberate conditions (Study 2: $F(1, 206) = 13.24$, $p < .001$, $\eta^2G = .061$; Study 3: $F(1, 198) = 8.92$, $p = .003$, $\eta^2G = .043$).

Table S58. Mean Accuracy by Condition for each Study (Conflict problems)

Study	Condition	Mean accuracy	SE
Study 1	Slow	14.98	3.93
Study 2	Fast	9.15	2.42
Study 2	Slow	26.28	4.01
Study 3	Fast	7.78	2.64
Study 3	Slow	21.82	3.66

Concerning argument evaluations, as can be seen in Figure S7 and S8, evaluation patterns when removing participants familiar with the classic bat-and-ball problems were highly similar to the patterns observed when including these participants (presented in the manuscript). This was the case in all three studies, for both fast-intuitive and slow-deliberate conditions. We detail the few contrasts where statistical results differed from the main analysis below.

In Study 1, the only differences with the results of the main analyses were that when removing participants familiar with the bat-and-ball problem, rating differences between correct responders' evaluations of incorrect-response arguments did not significantly differ from one another (all $p > .05$). Though patterns remain descriptively close to that of the main analysis, incorrect arguments with explicit heuristic justifications were no longer rated significantly higher than other incorrect arguments by the remaining participants who correctly solved the problem. One possible explanation for these differences turning non-significant is that removing participants familiar with the classic bat-and-ball problem left only 31 correctly solved conflict problems for analysis, which may constitute too little power to detect small difference in ratings. All other differences identified as significant in the main analysis remained significant in this control analysis.

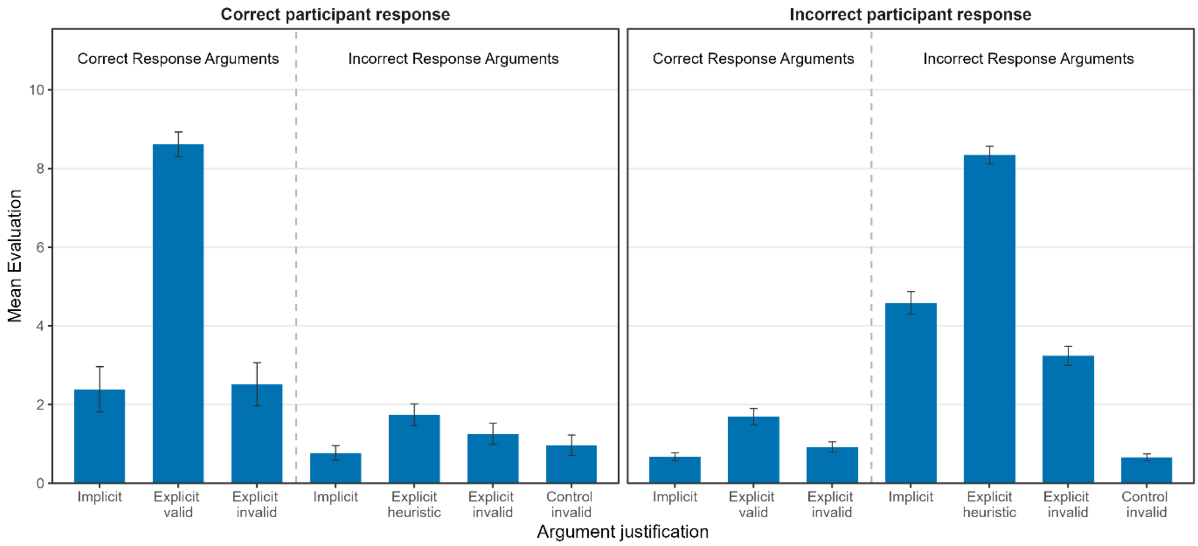

Fig. S9. Mean evaluations of arguments on conflict problems after correct (left panel) and incorrect (right panel) participant response, by argument response and justification, after removing participants familiar with the classic bat-and-ball problem (Study 1). Error bars are standard errors of the mean.

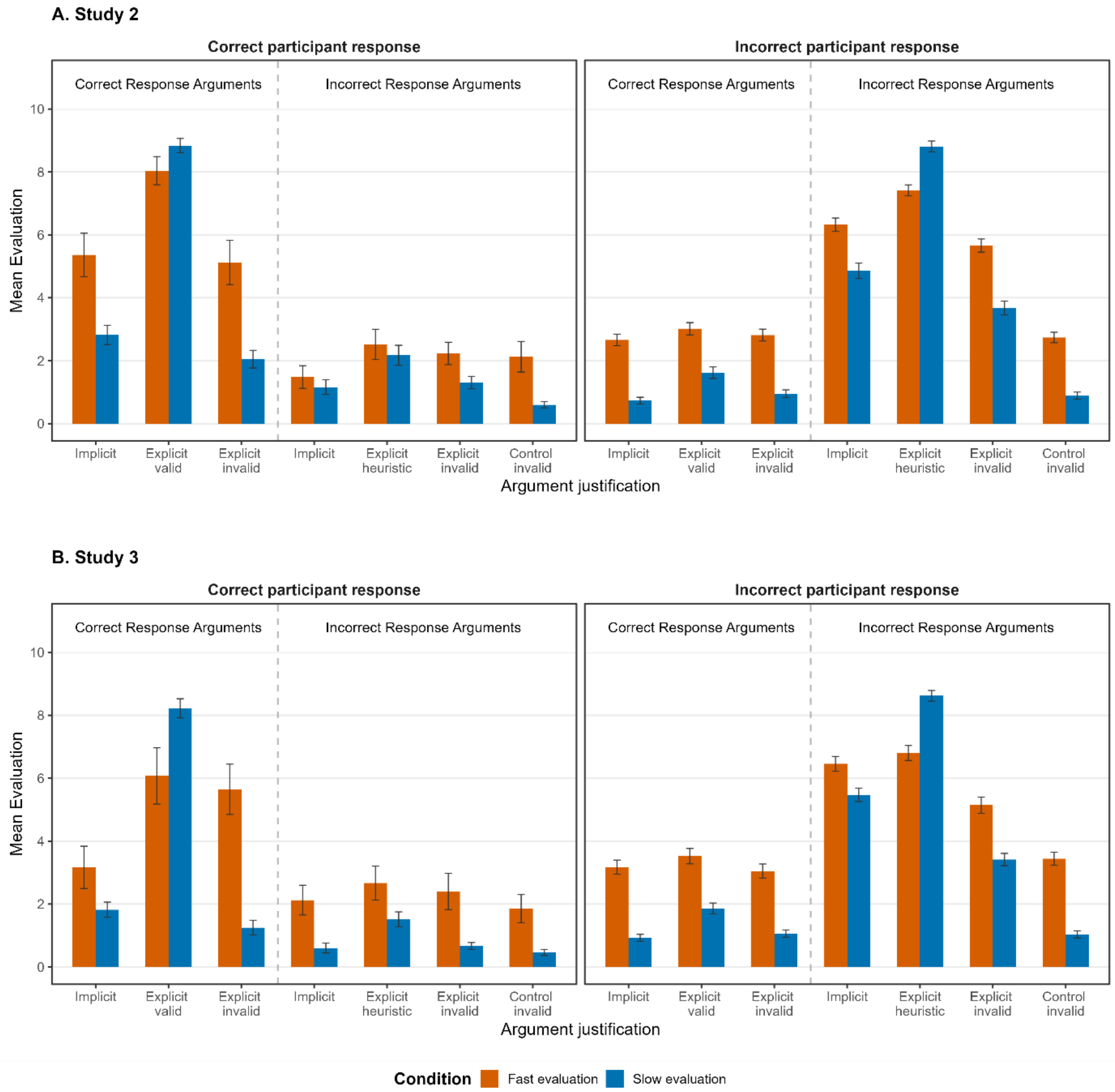

Fig.S10. Mean evaluations of arguments on conflict problems by condition in Study 2 (A) and Study 3 (B), separated by participant response to the problem and argument response correctness, after removing participants familiar with the classic bat-and-ball problem. Error bars are standard errors of the mean.

In Study 2, all differences identified as significant in the main analyses were also found significant in the control analysis without participants familiar with the classic bat-and-ball problem. This was the cas across correct and incorrect responders' evaluations, and in both fast-intuitive and slow-deliberate conditions.

In Study 3, concerning the slow-deliberate condition, patterns and statistical significance results for correct responders' evaluations of correct-response arguments were the same as those identified in the main analyses. For incorrect-response arguments however, while evaluations patterns were close to that of the main analysis, differences turned non-significant between correct responders' ratings of explicit heuristic arguments and that of both implicit heuristic ($t(35) = 2.62$, $p = .078$) and explicit invalid ones ($t(35) = 2.64$, $p = .074$). All evaluations from incorrect responders showed the same patterns and statistical significance as those reported in the main analysis. Concerning the fast-intuitive condition, the difference between correct responders' evaluations of implicit and explicit valid correct arguments turned non-significant ($p = .188$). For incorrect-response arguments, differences between implicit arguments and both explicit valid and explicit invalid arguments turned non-significant ($p = .492$ and $p = .428$ respectively). All evaluations from incorrect responders showed the same patterns and statistical significance results as those identified in the main analysis.

In addition, we also run a control analysis in which we included the ratings from participants who failed to correctly identify the matched matrix in the fast-intuitive conditions of Studies 2 and 3 but answered the rating question within the response window. Figure S9 shows the results when including these ratings in fast-intuitive evaluations. As can be seen in the figure, rating patterns were practically identical to those of the main analysis, in which fast-intuitive trials where excluded when participants failed to correctly recall the presented matrix. All statistical significance results were identical to those of the main analysis in Study 2. In Study 3, the only difference was that when including the failed recall trials, the difference between incorrect responders' ratings of implicit and explicit invalid incorrect arguments became only marginally significant ($t(209) = 2.56$, $p = .067$). Apart from this contrast, all statistical significance results were identical to those of the main analysis.

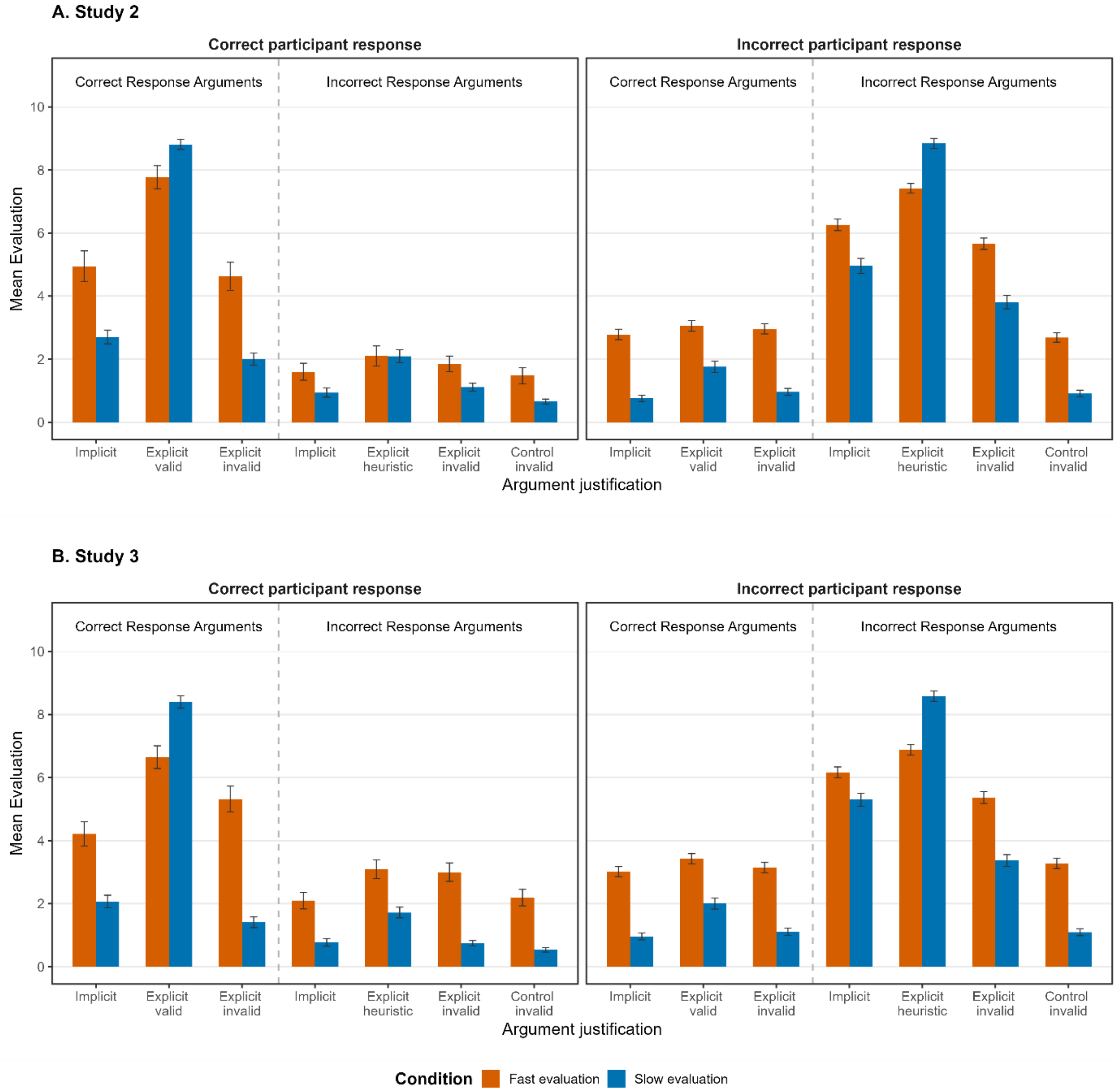

Fig.S11. Mean evaluations of arguments on conflict problems by condition in Study 2 (A) and Study 3 (B), separated by participant response to the problem and argument response correctness, when including ratings from participants who failed to identify the correct matrix in fast-intuitive conditions. Error bars are standard errors of the mean.

E. Additional analyses

Although consistency in response accuracy across the different conflict problems was relatively high, participants did not always respond consistently across problems, sometimes contributing both correct and incorrect trials. In our main analysis of argument evaluation in conflict problems, our correct- and incorrect-response split analyses therefore distinguish trials rather than two distinct groups of participants. Accordingly, these results should not be interpreted as comparing perfectly stable groups of “correct” and “incorrect” reasoners.

As noted by one reviewer, it would thus be possible that restricting the comparison to participants who responded consistently correctly or consistently incorrectly across problems would yield different patterns. Therefore, we report below an additional analysis including only participants who responded consistently correctly or consistently incorrectly across problems. As the interested reader will observe, the patterns were highly similar with that of the main analysis. Differences in significance from the main analysis are italicized and underligned.

Study 1

In Study 1, 24 participants always responded correctly to conflict problems, 63 always responded incorrectly, and 15 provided a mix of correct and incorrect responses.

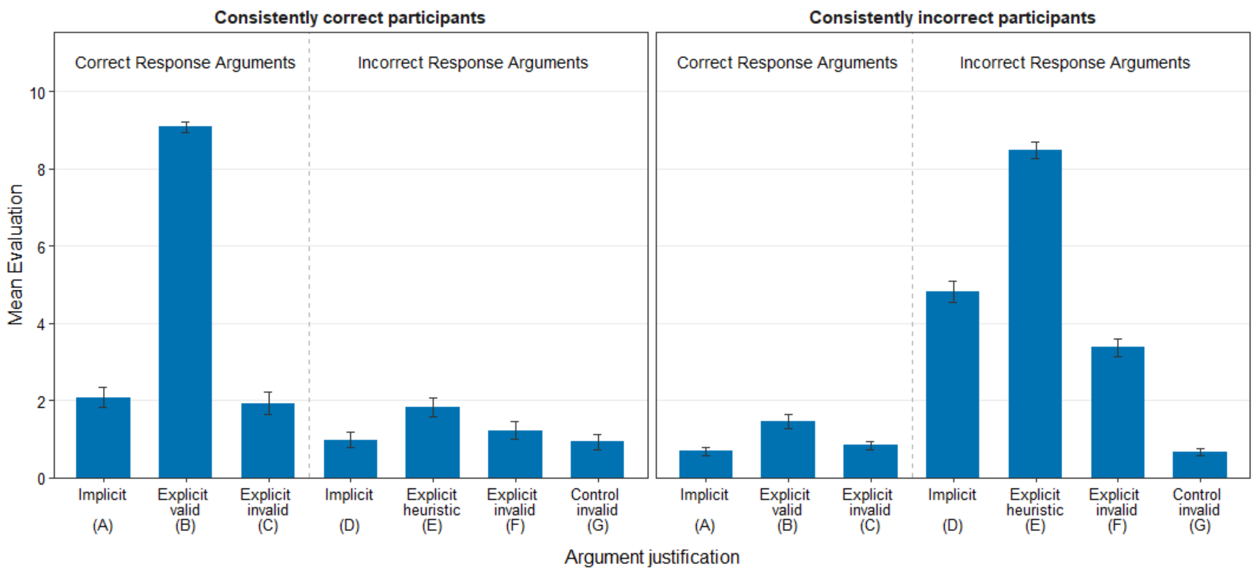

Fig.S12. Mean evaluation of arguments on conflict problems by condition in Study 1 for participants who were fully consistent across conflict trials (Left panel = always correct, Right panel = always incorrect). Error bars are standard errors of the mean.

Always Correct responders:

Table S59. ANOVA for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.874	43.093	3.34	128.084	0.768	<.001

Table S60. Post-hoc comparisons for correct arguments

contrast	estimate	SE	df	t.ratio	p.value
A - B	-7.000	0.514	23	-13.621	<.001
A - C	0.153	0.448	23	0.341	1.000
B - C	7.153	0.564	23	12.692	<.001

Table S61. ANOVA for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.991	45.795	0.75	8.233	0.039	<.001

Table S62. Post-hoc comparisons for incorrect arguments

contrast	estimate	SE	df	t.ratio	p.value
D - E	-0.847	0.257	23	-3.297	0.019
D - F	-0.236	0.191	23	-1.237	1.000
D - G	0.056	0.144	23	0.385	1.000
E - F	0.611	0.218	23	2.808	*0.060*
E - G	0.903	0.253	23	3.563	0.010
F - G	0.292	0.119	23	2.446	0.135

Always Incorrect responders:

Table S63. ANOVA for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.685	104.483	1.262	9.912	0.049	<.001

Table S64. Post-hoc comparisons for correct arguments

contrast	estimate	SE	df	t.ratio	p.value
A - B	-0.772	0.210	62	-3.672	0.002
A - C	-0.153	0.140	62	-1.092	0.837
B - C	0.619	0.193	62	3.207	0.006

Table S65. ANOVA for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.772	171.874	6.303	114.58	0.507	<.001

Table S66. Post-hoc comparisons for incorrect arguments

contrast	estimate	SE	df	t.ratio	p.value
D - E	-3.661	0.461	62	-7.936	<.001
D - F	1.450	0.446	62	3.253	0.011
D - G	4.148	0.481	62	8.629	<.001
E - F	5.111	0.429	62	11.923	<.001
E - G	7.810	0.379	62	20.622	<.001
F - G	2.698	0.374	62	7.222	<.001

Study 2

In the fast-intuitive condition of Study 2, 12 participants always responded correctly to conflict problems, 95 always responded incorrectly, and 18 provided a mix of correct and incorrect responses.

In the slow-deliberate condition of Study 2, 39 participants always responded correctly to conflict problems, 77 always responded incorrectly, and 17 provided a mix of correct and incorrect responses.

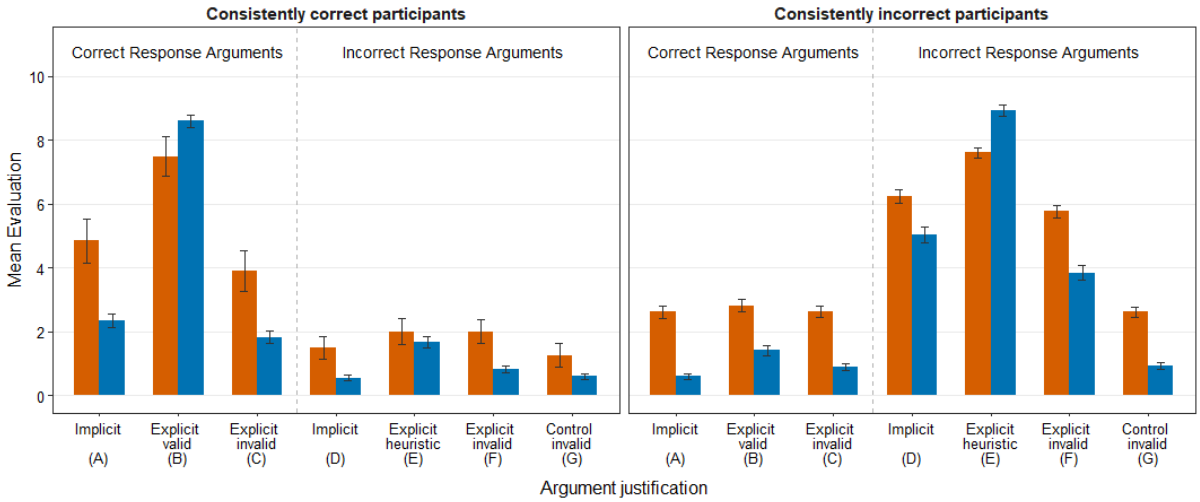

Fig.S13. Mean evaluations of arguments on conflict problems by condition in Study 2 for participants who were fully consistent across conflict trials (Left panel = always correct, Right panel = always incorrect). Error bars are standard errors of the mean.

Always Correct responders:

Table S67. ANOVA on Fast responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.647	18.117	5.847	11.602	0.273	<.001

Table S68. ANOVA on Slow responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.946	107.03	3.127	264.296	0.688	<.001

Table S69. Post-hoc pairwise comparisons for correct response arguments among correct responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
A - B	Fast	-3.250	0.762	49	-4.264	-0.597	<.001
A - C	Fast	0.833	0.715	49	1.165	0.163	0.749
B - C	Fast	4.083	0.738	49	5.534	0.775	<.001
A - B	Slow	-6.265	0.423	49	-14.817	-2.075	<.001
A - C	Slow	0.530	0.397	49	1.336	0.187	0.564
B - C	Slow	6.795	0.409	49	16.602	2.325	<.001

Table S70. ANOVA on Fast responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.608	17.69	3.691	0.468	0.015	0.593

Table S71. ANOVA on Slow responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.712	65.046	1.504	12.354	0.126	<.001

Table S72. Post-hoc pairwise comparisons on incorrect response arguments among always correct responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
D - E	Fast	-0.611	0.514	49	-1.189	-0.167	1.000
D - F	Fast	-0.500	0.307	49	-1.630	-0.228	0.657
D - G	Fast	-0.208	0.450	49	-0.463	-0.065	1.000
E - F	Fast	0.111	0.369	49	0.301	0.042	1.000
E - G	Fast	0.403	0.516	49	0.781	0.109	1.000
F - G	Fast	0.292	0.383	49	0.761	0.107	1.000

contrast	condition	estimate	SE	df	t.ratio	d	p.value
D - E	Slow	-1.128	0.285	49	-3.958	-0.554	0.001
D - F	Slow	-0.282	0.170	49	-1.657	-0.232	0.623
D - G	Slow	-0.060	0.250	49	-0.240	-0.034	1.000
E - F	Slow	0.846	0.205	49	4.133	0.579	<.001
E - G	Slow	1.068	0.286	49	3.735	0.523	0.003
F - G	Slow	0.222	0.213	49	1.045	0.146	*1.000*

Table S73. ANOVA: Correct vs Incorrect response arguments - Fast evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	11	2.879	24.448	0.442	<.001

Table S74. ANOVA: Correct vs Incorrect response arguments - Slow evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	38	1.196	182.71	0.695	<.001

Always Incorrect responders:

Table S75. ANOVA on Fast responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.91	168.086	1.386	0.662	0.001	0.510

Table S76. ANOVA on Slow responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.676	127.381	1.393	11.139	0.039	<.001

Table S77. Post-hoc pairwise comparisons for correct response arguments among incorrect responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
A - B	Fast	-0.199	0.188	164	-1.057	-0.081	0.876
A - C	Fast	-0.097	0.164	164	-0.594	-0.045	1.000
B - C	Fast	0.101	0.149	164	0.679	0.052	1.000
A - B	Slow	-0.810	0.202	164	-4.009	-0.306	<.001
A - C	Slow	-0.281	0.176	164	-1.595	-0.122	0.338
B - C	Slow	0.528	0.160	164	3.299	0.252	0.004

Table S78. ANOVA on Fast responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.507	225.661	5.83	89.347	0.362	<.001

Table S79. ANOVA on Slow responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.6	197.581	7.149	136.912	0.518	<.001

Table S80. Post-hoc pairwise comparisons for incorrect response arguments among always incorrect responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
D - E	Fast	-1.388	0.335	166	-4.141	-0.316	<.001
D - F	Fast	0.502	0.338	166	1.483	0.113	0.840
D - G	Fast	3.777	0.420	166	8.986	0.685	<.001
E - F	Fast	1.890	0.314	166	6.014	0.459	<.001
E - G	Fast	5.165	0.335	166	15.403	1.175	<.001
F - G	Fast	3.275	0.328	166	9.981	0.761	<.001
D - E	Slow	-3.905	0.364	166	-10.714	-0.817	<.001
D - F	Slow	1.186	0.368	166	3.224	0.246	0.009
D - G	Slow	4.108	0.457	166	8.992	0.686	<.001
E - F	Slow	5.091	0.342	166	14.899	1.136	<.001
E - G	Slow	8.013	0.365	166	21.983	1.676	<.001
F - G	Slow	2.922	0.357	166	8.192	0.625	<.001

Table S81. ANOVA: Correct vs Incorrect response arguments - Fast evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	94	2.718	131.756	0.296	<.001

Table S82. ANOVA: Correct vs Incorrect response arguments - Slow evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	76	2.023	263.64	0.575	<.001

Study 3

In the fast-intuitive condition of Study 3, 13 participants always responded correctly to conflict problems, 100 always responded incorrectly, and 10 provided a mix of correct and incorrect responses.

In the slow-deliberate condition of Study 3, 34 participants always responded correctly to conflict problems, 85 always responded incorrectly, and 19 provided a mix of correct and incorrect responses.

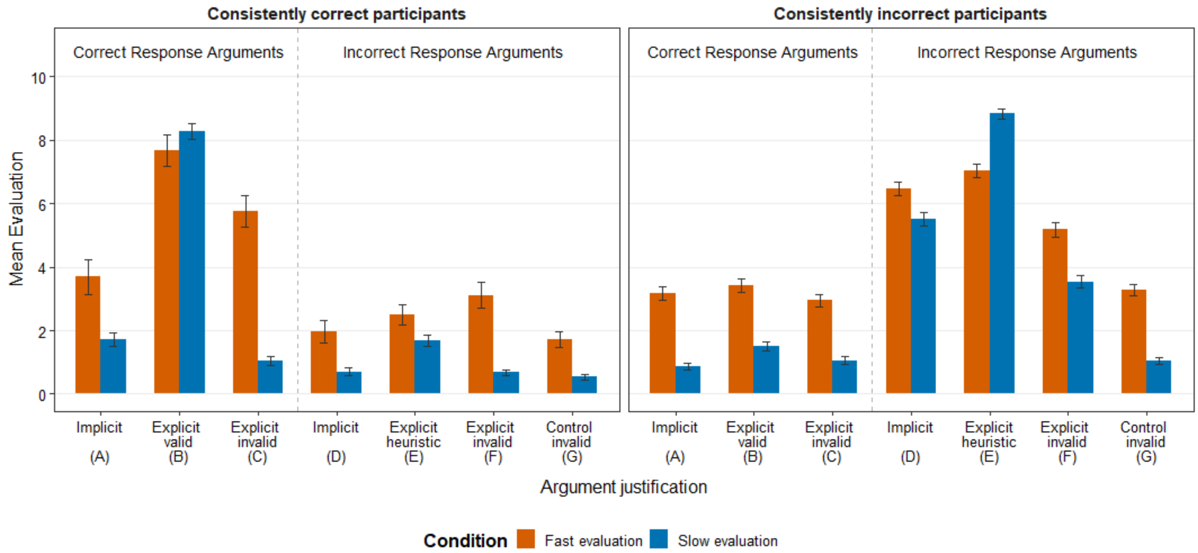

Fig.S14. Mean evaluations of arguments on conflict problems by condition in Study 3 for participants who were fully consistent across conflict trials (Left panel = always correct, Right panel = always incorrect). Error bars are standard errors of the mean.

Always Correct responders:

Table S83. ANOVA on Fast responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.119	8.949	2.995	16.991	0.231	0.002

Table S84. ANOVA on Slow responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.707	56.34	3.105	204.859	0.752	<.001

Table S85. Post-hoc pairwise comparisons for correct response arguments among correct responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
A - B	Fast	-3.556	0.862	41	-4.127	-0.602	<.001
A - C	Fast	-1.704	0.563	41	-3.028	-0.442	*0.013*
B - C	Fast	1.852	0.762	41	2.429	0.354	0.059
A - B	Slow	-6.559	0.443	41	-14.796	-2.158	<.001
A - C	Slow	0.676	0.289	41	2.337	0.341	0.073
B - C	Slow	7.235	0.392	41	18.446	2.691	<.001

Table S86. ANOVA on Fast responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.447	26.916	1.062	7.384	0.136	0.002

Table S87. ANOVA on Slow responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.972	65.065	1.22	11.813	0.138	<.001

Table S88. Post-hoc pairwise comparisons for incorrect response arguments among correct responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
D - E	Fast	-1.042	0.449	44	-2.321	-0.339	0.150
D - F	Fast	-1.458	0.344	44	-4.239	-0.618	<.001
D - G	Fast	-0.042	0.319	44	-0.131	-0.019	1.000
E - F	Fast	-0.417	0.421	44	-0.989	-0.144	1.000
E - G	Fast	1.000	0.392	44	2.552	0.372	*0.085*
F - G	Fast	1.417	0.257	44	5.511	0.804	<.001
D - E	Slow	-0.980	0.267	44	-3.677	-0.536	0.004
D - F	Slow	0.029	0.204	44	0.144	0.021	1.000
D - G	Slow	0.167	0.189	44	0.881	0.128	1.000
E - F	Slow	1.010	0.250	44	4.036	0.589	0.001
E - G	Slow	1.147	0.233	44	4.928	0.719	<.001
F - G	Slow	0.137	0.153	44	0.899	0.131	1.000

Table S89. ANOVA: Correct vs Incorrect response arguments - Fast evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	12	2.26	29.277	0.442	<.001

Table S90. ANOVA: Correct vs Incorrect response arguments - Slow evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	33	0.769	170.246	0.605	<.001

Always Incorrect responders:

Table S91.ANOVA on Fast responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
1.985	164.796	1.98	2.368	0.005	0.097	1.985

Table S92. ANOVA on Slow responses for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	1.649	138.524	1.19	9.292	0.024	<.001

Table S93. Post-hoc pairwise comparisons for correct response arguments among incorrect responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
A - B	Fast	-0.216	0.196	167	-1.103	-0.081	0.815
A - C	Fast	0.254	0.179	167	1.421	0.105	0.472
B - C	Fast	0.470	0.186	167	2.527	0.186	*0.037*
A - B	Slow	-0.639	0.195	167	-3.280	-0.242	0.004
A - C	Slow	-0.196	0.178	167	-1.104	-0.081	0.814
B - C	Slow	0.443	0.185	167	2.395	0.177	*0.053*

Table S94. ANOVA on Fast responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.266	174.458	7.227	41.161	0.235	<.001

Table S95. ANOVA on Slow responses for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
argument	2.576	216.381	6.438	166.246	0.545	<.001

Table S96. Post-hoc pairwise comparisons for incorrect response arguments among incorrect responders

contrast	condition	estimate	SE	df	t.ratio	d	p.value
D - E	Fast	-0.415	0.328	161	-1.264	-0.093	1.000
D - F	Fast	1.521	0.368	161	4.132	0.305	<.001
D - G	Fast	3.327	0.432	161	7.696	0.567	<.001
E - F	Fast	1.936	0.369	161	5.240	0.386	<.001
E - G	Fast	3.741	0.394	161	9.495	0.700	<.001
F - G	Fast	1.806	0.352	161	5.135	0.379	<.001
D - E	Slow	-3.333	0.314	161	-10.613	-0.782	<.001
D - F	Slow	1.965	0.353	161	5.571	0.411	<.001
D - G	Slow	4.455	0.414	161	10.757	0.793	<.001
E - F	Slow	5.298	0.354	161	14.971	1.104	<.001
E - G	Slow	7.788	0.377	161	20.633	1.521	<.001
F - G	Slow	2.490	0.337	161	7.394	0.545	<.001

Table S97. ANOVA: Correct vs Incorrect response arguments - Fast evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	98	3.084	76.417	0.22	<.001

Table S98. ANOVA: Correct vs Incorrect response arguments - Slow evaluations

	num Df	den Df	MSE	F	ges	Pr(>F)
argument_group	1	84	1.77	310.007	0.568	<.001

Contrast of Fast and Slow ratings

Study 2

Always Correct responders

Table S99. ANOVA for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
response_stage	1.000	49.000	7.898	3.558	0.038	0.065
argument	1.989	97.474	3.292	98.439	0.477	<.001
response_stage:argument	1.989	97.474	3.292	7.724	0.067	<.001

Table S100. ANOVA for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
response_stage	1.000	49.000	4.851	5.889	0.067	0.019
argument	2.383	116.777	1.397	4.860	0.039	0.006
response_stage:argument	2.383	116.777	1.397	0.913	0.008	0.419

Always Incorrect responders

Table S101. ANOVA for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
response_stage	1.000	164.000	11.801	31.934	0.138	<.001
argument	1.862	305.335	1.344	8.554	0.009	<.001
response_stage:argument	1.862	305.335	1.344	3.242	0.003	0.044

Table S102. ANOVA for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
response_stage	1.000	166.000	11.643	13.585	0.033	<.001
argument	2.698	447.916	6.091	224.475	0.442	<.001
response_stage:argument	2.698	447.916	6.091	16.110	0.054	<.001

Study 3

Always Correct responders

Table S103. ANOVA for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
response_stage	1.000	41.000	8.433	8.192	0.112	0.007
argument	1.658	67.984	2.968	89.633	0.446	<.001
response_stage:argument	1.658	67.984	2.968	21.050	0.159	<.001

Table S104. ANOVA for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
response_stage	1.000	44.000	4.342	17.314	0.201	<.001
argument	2.443	107.489	1.005	12.218	0.091	<.001
response_stage:argument	2.443	107.489	1.005	5.379	0.042	0.003

Always Incorrect responders

Table S105. ANOVA for correct arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
response_stage	1.000	167.000	11.685	42.486	0.169	<.001
argument	1.977	330.179	1.487	7.516	0.009	<.001
response_stage:argument	1.977	330.179	1.487	1.830	0.002	*0.162*

Table S106. ANOVA for incorrect arguments

	num Df	den Df	MSE	F	ges	Pr(>F)
response_stage	1.000	161.000	11.463	6.571	0.016	0.011
argument	2.727	438.976	6.045	179.196	0.396	<.001
response_stage:argument	2.727	438.976	6.045	23.527	0.079	<.001

Table S107. Post-hoc pairwise comparisons of fast vs slow ratings

Study	Participant response	Argument	Estimate (Fast-Slow)	SE	df	t.ratio	d	p.value
Study 2	Always Incorrect	Correct implicit	2.009	0.318	164	6.323	1.025	<.001
		Correct explicit valid	1.398	0.371	164	3.768	0.589	<.001
		Correct explicit invalid	1.825	0.328	164	5.560	0.881	<.001
		Incorrect (heuristic) implicit	1.349	0.514	166	2.625	0.400	0.009
		Incorrect (heuristic) explicit valid	-1.168	0.345	166	-3.383	-0.551	<.001
		Incorrect (heuristic) explicit invalid	2.033	0.437	166	4.648	0.702	<.001
		Incorrect control invalid	1.680	0.314	166	5.350	0.853	<.001
	Always Correct	Correct implicit	2.116	0.833	49	2.540	0.839	0.014
		Correct explicit valid	-0.899	0.669	49	-1.344	-0.444	0.185
		Correct explicit invalid	1.813	0.658	49	2.755	0.910	0.008
		Incorrect (heuristic) implicit	0.920	0.379	49	2.426	0.801	*0.019*
		Incorrect (heuristic) explicit valid	0.403	0.591	49	0.682	0.225	0.499
		Incorrect (heuristic) explicit invalid	1.138	0.430	49	2.649	0.874	*0.011*
		Incorrect control invalid	1.068	0.463	49	2.309	0.762	0.025
Study 3	Always Incorrect	Correct implicit	2.270	0.325	167	6.977	1.092	<.001
		Correct explicit valid	1.847	0.362	167	5.103	0.802	*<.001*
		Correct explicit invalid	1.820	0.331	167	5.506	0.860	<.001
		Incorrect (heuristic) implicit	1.017	0.479	161	2.122	0.342	0.035
		Incorrect (heuristic) explicit valid	-1.902	0.378	161	-5.031	-0.759	<.001
		Incorrect (heuristic) explicit invalid	1.461	0.451	161	3.237	0.552	0.001

Table S107. Post-hoc pairwise comparisons of fast vs slow ratings

Study	Participant response	Argument	Estimate (Fast-Slow)	SE	df	t.ratio	d	p.value
		Incorrect control invalid	2.145	0.333	161	6.443	0.971	<.001
	Always Correct	Correct implicit	2.007	0.865	41	2.320	0.863	0.025
		Correct explicit valid	-0.997	0.872	41	-1.143	-0.467	*0.260*
		Correct explicit invalid	4.387	0.606	41	7.236	2.853	<.001
		Incorrect (heuristic) implicit	1.016	0.432	44	2.354	0.753	0.023
		Incorrect (heuristic) explicit valid	1.078	0.549	44	1.964	0.548	*0.056*
		Incorrect (heuristic) explicit invalid	2.504	0.413	44	6.068	2.037	<.001
		Incorrect control invalid	1.225	0.329	44	3.720	1.304	<.001