Cognition 240 (2023) 105380

Contents lists available at ScienceDirect

Cognition

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

The formation and revision of intuitions☆

Andrew Meyer [a,*], Shane Frederick [b,*]

[a] The Chinese University of Hong Kong (CUHK) Business School, Department of Marketing, 12 Chak Cheung St, Shatin, N.T., Hong Kong SAR
[b] Yale School of Management, Marketing Department, 165 Whitney Ave, New Haven, CT 06511, USA

ARTICLE INFO

Keywords:
Dual systems theory
Cognitive reflection test

ABSTRACT

This paper presents 59 new studies (N = 72,310) which focus primarily on the "bat and ball problem." It documents our attempts to understand the determinants of the erroneous intuition, our exploration of ways to stimulate reflection, and our discovery that the erroneous intuition often survives whatever further reflection can be induced. Our investigation helps inform conceptions of dual process models, as "system 1" processes often appear to override or corrupt "system 2" processes. Many choose to uphold their intuition, even when directly confronted with simple arithmetic that contradicts it – especially if the intuition is *approximately* correct.

Mental operations range from rapid, effortless, perceptual impressions (recognizing a face) to more deliberate computations that one must choose to execute (algebra). Sometimes, operations that are effortful initially (eleven minus three) become nearly automatic later.

Research examining the ease or difficulty of mental operations often goes under the label of dual systems or dual processes. In the framework advanced by Kahneman and Frederick (2002, 2005), a fast and intuitive system proposes initial answers which a slower, more reflective system scrutinizes and then accepts, rejects, or revises. Others have produced similar frameworks (Sloman, 1996; Stanovich & West, 2000).

We assume that exposure to any stimulus will initiate *at least* one cognitive process, which takes some time to execute. The output of that process may be sufficient to permit a response. In other cases, the stimulus evokes a second cognitive process which may be initiated concurrently with the first, somewhat later (as shown in Fig. 1), or after the first process has yielded some output needed to initiate a subsequent operation. If a second process is initiated, we assume only that it concludes after the first, and that its output may either affirm or compete with the output of the first process for control of the overt response.[1]

In the most discussed examples in dual process research, subsequent considerations conflict with an initial impression. Consider "Linda," who Tversky and Kahneman (1983) described as having opposed nuclear power and taken an interest in issues of discrimination. Subjects must decide whether she is more likely to be:

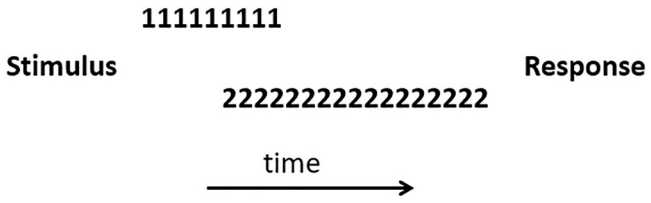


Fig. 1. Generic dual process model

(a) a bank teller
(b) a bank teller who is active in the feminist movement

Given Linda's description, most readily conceptualize her as a

☆ We thank Yigal Attali, Jon Baron, Maya Bar-Hillel, Christopher Chabris, Ripley Chance, Zoë Chance, Phil Cortlett, Wim De Neys, Reid Hastie, Daniel Kahneman, Gideon Keren, Jin Kim, Jack Klinger, Asher Lawson, Steve Malliaris and Robert Spunt for discussion and feedback.

* Corresponding authors.
E-mail addresses: andrewmeyer@cuhk.edu.hk (A. Meyer), shane.frederick@yale.edu (S. Frederick).

[1] Under this conception, *parallel-competitive* models (Sloman, 1996) and *default-interventionist* models (Kahneman & Frederick, 2002, 2005) are distinguished "only" with respect to whether the second process is initiated before or after the first has concluded. It isn't always clear *what* initiates a second process, absent, say, an explicit entreaty to verify that the provided answer is correct (see Pennycook et al., 2015).

https://doi.org/10.1016/j.cognition.2023.105380
Received 4 February 2022; Received in revised form 11 January 2023; Accepted 16 January 2023

feminist, and are eager to express that inference. However, those who scrutinize this intuition may recognize that the set of bank tellers encompasses feminist bank tellers, and thereby change their answer from (b) to (a). We'd consider that to be the reflective response, as it is emitted later and associated with superior reasoning on other tasks.

Analogously, consider the question below:

Were the 9/11 hijackers cowards? Yes No

The intuitive response here is "Yes," because most are eager to attach a negative label to a negatively evaluated target. Once again, however, those who think "harder" may question the suitability of that label. Accordingly, "No" responses are produced more slowly and also betoken superior reasoning abilities (see Appendix A).

In the questions above, the intuitions (feminists care about discrimination, hijackers are bad) exist apart from the stimulus. In other cases, the intuition may emerge from an operation on elements within the stimulus. Consider the "bat & ball" problem below.[2]

A bat and a ball cost $1.10 in total.
The bat costs $1.00 more than the ball.

How much does the ball cost? ___ cents

The first sentence references two objects and specifies their sum ($1.10). The second includes the words "more than," inviting respondents to subtract something from that sum, with the only remaining number ($1.00) providing an attractive candidate. The subtraction yields a 10-cent ball, which is the modal response. With a 10-cent ball, the problem's two requirements cannot be mutually satisfied. If the two prices sum to $1.10, the bat must cost $1.00 (which is only 90 cents more than a 10-cent ball). If the two prices differ by $1.00, the bat, itself must cost $1.10 (and the two prices would sum to $1.20).

The bat and ball problem is often used to illustrate a dual process model of cognition, in which a fast "system" provides tentative answers that a slower "system" inspects and (only) sometimes revises. This conception is supported by observations that the intuitive answer is (a) initially considered by many who ultimately respond correctly (Frederick, 2005; Szaszi, Szollosi, Palfi, & Aczel, 2017; Travers, Rolison, & Feeney, 2016), (b) more common under mnemonic load or time constraints (Borghans, Meijers, & Ter Weel, 2008; Johnson, Tubau, & De Neys, 2016) and (c) produced more quickly than the correct answer, despite the superior computational abilities of those who respond correctly (see Appendix B).

Kahneman and Frederick (2002, 2005) proposed that judgmental errors raise questions about both the production of erroneous intuitions ("System 1" questions) and the factors that facilitate or inhibit their detection and revision ("System 2" questions). We will retain this schema to organize our discussion, but will also raise objections to it along the way.

We first show that there are multiple routes to the 'intuitive' response, as distinctions can be made even among those who say 10; some do so because they misread the question, whereas others appear to just subtract the smaller number from the larger one, with little regard for the words in which those numbers are embedded. Such results are not easily situated within dual-system frameworks, which usually emphasize differences in the degree to which intuitions are scrutinized, but neglect differences in the sophistication or complexity of operations that led to the so-called intuition.

The second part of our paper further chafes dual system models. In contrast with the common assumption that reflection will reject faulty intuitions, we find that intuitions typically survive whatever kinds of reflection we can induce. Executing the intuitive operation appears to inhibit or impair the reasoning processes needed to detect the error. For example, upon concluding that the bat costs $1.00 and the ball costs $0.10, many respondents will explicitly affirm that those two values differ by $1.00.

Although our attempts to induce reflection only slightly raised performance for the standard problem, we later show that performance improves dramatically for variants of the problem in which the heuristic operation yields results that more radically violate the stipulated constraints. We propose the notion of an "approximate checker" which pardons small errors but not large ones.

We conclude by situating these findings within a broader discussion of dual system theories and propose ways of distinguishing the willingness to think from the ability to think.

1. Forming intuitions

Kahneman and Frederick (2002, 2005) used the term *attribute substitution* to describe situations in which respondents unwittingly answer a simpler version of the question they encounter. With that concept in mind, consider a "lite" version of the bat and ball problem.

A bat and a ball cost $1.10 in total.
The bat costs $1.00.

How much does the ball cost? ___ cents

Of course, 10 cents *is* the correct answer to *this* question. If you find yourself stopping here to ponder how it differs from the original, you can appreciate how readily it might be substituted.[3] Further evidence for this substitution is revealed when respondents must specify the price of *both* items.

A bat and a ball cost $1.10 in total.
The bat costs $1.00 more than the ball.

How much does the ball cost? ____
How much does the bat cost? ____

Among 196 mTurk participants, 121 made the common error, concluding that the ball cost 10 cents,[4] and all but two of them concluded that the bat cost $1.00. In other words, nearly everyone who missed the problem generated prices which satisfied its *first* constraint (summing to $1.10) but violated its *second* (differing by $1.00). This result suggests that respondents were substituting the simpler "lite" version (where the bat costs $1.00) for the actual question (where the bat costs $1.00 *more than* the ball).

We further tested the hypothesized substitution by asking 615 MTurkers to reproduce the problem from memory. Among those who

[2] This is the best known of the three items comprising the "Cognitive Reflection Test" or CRT proposed by Frederick (2005).

[3] The lite version is not only easier, but more typical. In his survey of mathematics textbooks, Mayer (1981) found that word problems which assign values to *variables* (as in the lite version) are six times more common than those which assign values to *relations* (as in the standard problem). Thus, much as people are prone to apply a solution strategy from a preceding problem to a subsequent problem (Luchins, 1942) they are more likely to apply solution strategies from problems they encounter more frequently.

[4] Widespread exposure to this problem on Amazon's Mechanical Turk is a well-known issue. Although repeated exposure has surprisingly little impact on the item's predictive validity (Bialek & Pennycook, 2018; Meyer, Zhou, & Frederick, 2018; Stagnaro, Pennycook, & Rand, 2018), it could still affect response processes. Accordingly, for all of our MTurk studies (and some of our other studies), we asked participants whether they had seen the problem before and excluded those who said they had. Thus, the referenced Ns in the paper refer to the *subset* of participants who were plausibly seeing the item for the first time. Appendix C reports demographics for all studies reported in the main text.

solved the problem, nobody misremembered it as the lite variant, but among those who made the 10-cent error, 23% did so. Although this is some evidence for the posited substitution, 61% of those who said 10 cents *could* recall the words that their answer implies they neglected – "*more than* the ball." (See Appendix D and Hoover & Healy, 2019.) Moreover, we found no effect of emphasizing the "neglected" detail, such as by bolding the words **more than the ball.** (We discuss these studies in Appendix E, though see Hoover & Healy, 2019; Mata, 2020; Mata, Ferreira, & Sherman, 2013 who each found effects using comparable manipulations on smaller samples.)

We initially regarded the posited substitution as the thoughtless error, but later learned that many respondents follow an even simpler strategy: subtracting the smaller number from the larger one. In the standard problem, these two strategies yield the same answer (10), but if one instead asks about the price of the *bat* (as below) substitution would yield the answer 100, whereas subtraction would yield the answer 10.[5]

A bat and a ball cost $110 in total.
The bat costs $100 more than the ball.

How much does the bat cost? ____

Among 1001 respondents on Google Consumer Surveys (hereafter GCS) who answered the "bat price" problem, the $10 "subtraction" response was nearly as common as the $100 "substitution" response (31% vs. 34%), was emitted much faster (23 s vs. 36 s),[6] and was associated with even shallower reasoning in other tasks (see Appendix F).[7]

This very simple and very fast subtraction "strategy" is also evident in the "*lite difference*" variant below:

A bat and a ball cost $110 in total.
The bat costs $100.

What is the difference in price between the bat and the ball? ____

Here, there is no opportunity to misinterpret the second number as the price of the bat, because it *is* the price of the bat. However, the second number can still be subtracted from the first, and many do that: among 1032 GCS respondents, 56% answered $10.[8]

These results create issues for dual process theories, by suggesting *three* "types" or "levels" of thought: (1) a super-fast subtraction strategy in which the smaller number is subtracted from the larger one, (2) a medium speed strategy involving the unwitting substitution of a similar, but simpler, question, and (3) a slow strategy of generating values that actually satisfy both of the stipulated constraints. This ternary classification may be accommodated by the dual system nomenclature if one regards the "two" so called "systems" as endpoints on some thought continuum, but such results still raise questions about the criteria used to position responses (or people) on that continuum. Should thinking "styles" or "levels" be characterized in terms of the overt response, *own* reaction time, average reaction time of *others* who produced that response, or from evidence that an initial thought was overridden (even if replaced by a new thought that was *also* incorrect)?

2. Maintaining intuitions

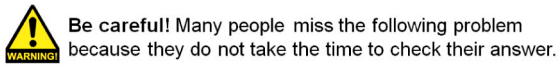
Whether resulting from subtraction or substitution, respondents are highly and often *maximally* confident that their $0.10 response is correct (see Appendix G). Why does the error remain hidden in plain sight? The constraint that the two prices differ by $1.00 is one of just three sentences, clearly stated, and sometimes even emphasized. Moreover, verifying the intuitive response requires nothing more than adding $1.00 and $0.10 to ensure that they sum to $1.10 (they do) and subtracting $0.10 from $1.00 to ensure that they differ by $1.00 (they don't). Since essentially everyone *can* perform these verification tests, the high error rate means that they aren't being performed or that respondents are drawing the wrong conclusion despite performing them.

If respondents aren't attempting to verify their answer, encouraging them to do so may help. We tested this in five studies involving a total of 3219 participants who were randomly assigned to either a control condition or to one of four warning conditions shown below. Two studies were administered to students who used paper and pencil. The rest were web-based surveys of a broader population.[9]

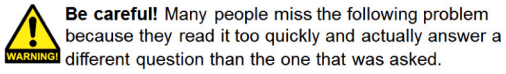
***SIMPLE* WARNING**

Be careful! Many people miss this problem.

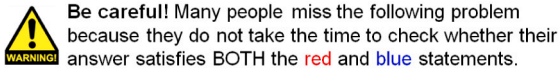
***COMPUTATION* WARNING**



***COMPREHENSION* WARNING**



***CONSTRAINT* WARNING**



The warnings improved performance, but not by much (see Table 1). This suggests that they failed to engage a checking process, or that the checking process was insufficient to remedy the error.[10] Others find similarly modest effects of asking respondents to reflect on initial responses, for the bat and ball problem (Bago & De Neys, 2019) and for other reasoning tasks (Lawson, Larrick, & Soll, 2020; Thompson, Turner, & Pennycook, 2011).

[5] To avoid the need to recode decimal errors, here we specified the prices as dollars rather than cents. We use both versions of the problem throughout this paper. Solution rates are about the same.

[6] Unless otherwise specified, response times are geometric means.

[7] The *bat price* problem has the advantage of partitioning subjects into three "tiers" of reasoning, rather than two. A control condition ($N = 1009$) revealed that it is solved at the same rate as the *standard* problem (20% vs. 19%).

[8] The *lite difference* problem is solved more often (32%) than the *bat price* problem (20%) or *standard* problem (19%), and more quickly (33 s vs. 55 s and 48 s). However, for all three problems, the $10 error is comparably fast (23 s, 23 s, and 25 s, respectively).

[9] Here, and in the next study, the problem was sometimes presented by itself and sometimes as the first item in the 3-item CRT (Frederick, 2005). In the Constraint Warning condition, the problem's first sentence "A bat and a ball cost $1.10 in total." was printed in red; and its second "The bat costs $1.00 more than the ball." was printed in blue.

[10] Although warnings have only modest effects on solution rates, they do increase time spent on the problem. We presume that this extra time was spent engaged in mental activity related to the problem, which one might reasonably call "checking." Nevertheless, as discussed in Appendix G, these checks not only failed to markedly improve performance, they also failed to reduce confidence in the erroneous intuition.

Table 1
Effects of various warnings

	Google	mTurk	UCLA[a]	eLab	Yale
	N = 2003	N = 238	N = 282	N = 454	N = 241
Control (none)	13 $_{23}$	34 $_{22}$	41 $_{-}$	43 $_{21}$	79 $_{-}$
Simple	23 $_{34}$	–	–	–	–
Computation	–	25 $_{33}$	49 $_{-}$	52 $_{37}$	84 $_{-}$
Comprehension	–	44 $_{31}$	58 $_{-}$	40 $_{39}$	91 $_{-}$
Constraint	–	38 $_{38}$	–	35 $_{38}$	–

Main script indicates percent correct. Subscript indicates seconds to respond.
[a] A special thank you here to Bob Spunt, who helped design this study and collected these data.

Since these warnings were ineffective, we next tried an even stronger manipulation by telling respondents that 10 cents is not the answer. We conducted eight such experiments, with a total of 7766 participants. In five studies (three online and two paper and pencil), participants were randomly assigned to either the *control* condition or to a *Hint* condition in which the words "HINT: 10 cents is not the answer" appeared next to the response blank.

A bat and a ball cost $1.10 in total.
The bat costs $1.00 more than the ball.

How much does the ball cost? $___
HINT: 10 cents is not the answer.

In three other studies (two online and one in-lab), we used a within-participant design in which the *Hint* was provided *after* the participant's initial response. In those studies, respondents could revise their initial (unhinted) response, and we recorded both their initial and final responses. The results of all eight studies are shown below in Table 2.

The hint that the answer *wasn't* 10 cents helped substantially, but, more notably, many – and sometimes most – still failed to solve the problem.[11] Though the bat and ball problem is often used to categorize people as reflective (those who say 5) or intuitive (those who say 10), these results suggest that the "intuitive" group can – and *should* – be further divided into the "careless" (who answer 10, but revise to 5 when told they are wrong) and the "hopeless" (who are unable or unwilling to compute the correct response, even when told that 10 is not the answer). The *careless* fit neatly into the dual process framework, but the *hopeless* do not, and they create problems for those using this item as a measure of reflection in non-elite populations. If the very thing that reflection selectively provides to those who have enough of it – realization that the answer *cannot* be 10 – is provided to *all* by telling them that the answer *is* not 10, and responses still vary, the problem must also be measuring other things.[12] As shown in Fig. M of Appendix M, the relative size of these three groups depends on the cognitive abilities of the populations being tested: Though many highly intelligent people get this problem wrong, nearly all of them are *careless*, whereas many others cannot solve it, even after being alerted to the common error, suggesting that it requires more effort than they are willing to expend or greater abilities than they possess. A recent study (Enke et al., 2021) implies the latter, as offering participants a full month's salary for solving the problem only modestly increased solution rates (from 35% to 48%).

Manipulations intended to raise solution rates may fail to do so in part because the operations producing the intuition disrupt or degrade the execution of subsequent operations needed to detect the error. Consider the results of the following study, in which 2010 GCS respondents were randomly assigned to one of two conditions.

MINUEND ABSENT
A bat costs $100 more than a ball

If you said the bat cost $100 and the ball costs $10, would your prices be correct? **18%** YES NO

MINUEND PRESENT
A bat and a ball cost $110 in total.
The bat costs $100 more than the ball.

If you said the bat cost $100 and the ball cost $10, would your prices be correct? 53% YES NO

When the heuristic operation is encouraged by supplying the 110 minuend (from which 100 might be subtracted) many more respondents erroneously affirm $100 and $10 as the correct prices, even though *both* conditions clearly stipulate that the bat costs $100 *more*, and even though rejecting that pair of prices requires no further mental effort: no need to determine the cost of either object.

A subsequent study reveals that once the outputs of the intuitive operation are expressed, they become even more recalcitrant to requests for further scrutiny. In that study, 124 passengers on a commuter ferry between Connecticut and Long Island were either told that a bat cost $1.00 and a ball cost $0.10, or were presented with the standard question, which required them to generate prices for each object. We then asked all participants whether the pair of prices (which they had either been *provided* with or *generated*) differed by $1.00. Among those *provided* with a $1.00 bat and $0.10 ball, only 6% said "Yes," whereas 76% of those who had *generated* those same two prices did so.[13] Once again, the heuristic operation (subtraction) appears to disrupt or degrade

Table 2
Effects of "Hint: 10 cents is not the answer."

	Between-subject experiments				
	Google N = 2635	eLab N = 562	mTurk N = 360	UCLA N = 551	Yale N = 275
Control	14 $_{71}$	31 $_{60}$	38 $_{54}$	42 $_{56}$	65 $_{31}$
Hint	34 $_{20}$	54 $_{12}$	65 $_{22}$	64 $_{6}$	82 $_{3}$
	Within-subject experiments				
		eLab N = 196	mTurk N = 3137		Yale N = 50
Control	–	26 $_{58}$	29 $_{63}$	–	50 $_{26}$
Hint	–	46 $_{19}$	51 $_{27}$	–	66 $_{6}$

Main script indicates percent correct. Subscript indicates percent responding 10 cents.

[11] These effects are much larger than those of similar hints administered after participants have already attempted multiple variants of the problem (Janssen, Raoelison, & de Neys, 2020), but somewhat smaller than removing the 10-cent lure from a set of response options which include the correct answer (Patel, Baker, & Scherer, 2019).

[12] An examination of Table 2's subscripts reveals that some of the "hopeless" may be better described as *stubborn*: they maintain their 10-cent response despite the hint that that answer is wrong. In our three within-subject studies, some participants seem to have assumed that we were repudiating the *form* of their 10-cent response rather than its *content*, as they modified their response from one form of 10 cents to another – such as rewriting a decimal response (0.1) as a whole number (10).

[13] Among the 57 participants in the *Generated* condition, 41 (or 72%) entered the two intuitive prices. Another 12 participants gave the correct pair of prices ($0.05 and $1.05) and all of them affirmed the $1.00 difference.

subsequent operations involving its output. (See Appendices H and I for further data and discussion.)

PROVIDED PRICES
A bat costs $1.00 and a ball costs $0.10.

	6%	
With those prices, does the bat cost $1.00 more than the ball?	YES	NO

GENERATED PRICES
A bat and a ball cost $1.10 in total.
The bat costs $1.00 more than the ball.
How much does the ball cost? $ 0.10
How much does the bat cost? $ 1.00

	76%	
Is your "bat" answer $1.00 more than your "ball" answer?	YES	NO

In stark contrast to the account that the 10-cent error indicates an unwillingness to check (Frederick, 2005; Kahneman & Frederick, 2002, 2005), the error survives at least some cursory version of the checking process that ought to expose it. Even when their attention is directed to the constraint specifying that prices differ by $1.00, most respondents nevertheless maintain that their $1.00 and $0.10 responses satisfy that constraint. This result has hallmarks of simultaneous contradictory belief (Sloman, 1996), because respondents who report that $1.00 and $0.10 differ by $1.00 obviously do not actually believe this. It is also akin to research on Wason's four card task showing that participants will rationalize their faulty selections, rather than change them (Beattie & Baron, 1988; Wason & Evans, 1974). It could also be considered as an *Einstellung* effect (Luchins, 1942), in which prior operations blind respondents to an important feature of the current task or as an illustration of confirmation bias, in which initial erroneous interpretations interfere with the processes needed to arrive at a correct interpretation (Bruner & Potter, 1964; Nickerson, 1998).

The durability of the intuition is further evidenced by the types of manipulations researchers have resorted to, such as providing respondents with arguments for *why* $5 is correct (Trouche, Sander, & Mercier, 2014) or testing whether *classrooms* of university students asked to discuss the problem can reach a correct consensus (Claidière, Trouche, & Mercier, 2017 showed, reassuringly, that they can).

Similarly, as shown below, we ran two studies on GCS in which we asked respondents to either *consider* the correct answer ($N = 2002$) or to simply *enter* it ($N = 1001$).

Consider $5
A bat and a ball cost $110 in total.
The bat costs $100 more than the ball.

How much does the ball cost?
Before responding, consider whether the answer could be $5.
$___

Enter $5
A bat and a ball cost $110 in total.
The bat costs $100 more than the ball.

How much does the ball cost?
The answer is $5.
Please enter the number 5 in the blank below.
$___

Asking respondents to *consider* the correct answer more than doubled solution rates, but only to 31%. Asking them to simply enter the correct answer worked better, as 77% did so, though, notably, the intuitive response emerged even here. See Appendix J for further data.

Of course, interpreting the results from such extreme manipulations as "solution rates" obviously distorts what it means to "solve" a problem. Moreover, the very existence of such manipulations (and their lack of *complete* efficacy) undermines a conclusion many draw from dual process theories of reasoning: that judgmental errors can be avoided *merely* by getting respondents to slow down and think harder.[14]

3. Revising intuitions

Historically, the bat and ball problem has been used to illustrate the laziness or inefficacy of corrective operations (Kahneman & Frederick, 2002, 2005). We endorse this view, but further suggest that the presence of the erroneous intuition prevents the correct conclusion from being drawn even when checks *are* attempted. An explicit affirmation that $1.00 and $0.10 differ by $1.00 suggests that the erroneous intuition corrupts any subsequent operations; it is not merely *acceded* to (see Risen, 2016), but *endorsed*.

There are, however, limits to this endorsement. Much as people can more quickly distinguish numbers that are further apart (Moyer & Landauer, 1967), their endorsement of the intuitive operation diminishes when it yields values that more strongly violate the problem's requirements. To illustrate this, we randomly assigned 10,044 GCS participants to one of ten conditions. The problem always began as usual: "A bat and a ball cost $110 in total…" but we varied the number specified in the second sentence: "The bat costs [X] more than the ball," with values ranging from $100 to $10. As the specified difference gets smaller, the intuitive operation (subtracting the difference from the total) yields values that more radically violate the problem's requirements. For example, if the specified difference is $40, the intuitive operation ($110 minus $40) yields a $70 ball, which would be more than half of the $110 total it is supposed to share with a more expensive item.

For each of those ten conditions, Fig. 2 shows the fraction who *Solve* the problem, who *Subtract* the difference from the total (the posited intuitive operation), or who give some *Other* answer. As the price difference between the bat and ball decreases, participants slow down (see Appendix K) and solution rates rise markedly – from 14% to 57%.[15] In other words, the results of the posited intuitive operation (subtraction) appear to receive greater scrutiny when yielding values that more radically violate the problem's requirements.[16]

To illustrate our "approximate checker" hypothesis, suppose that someone charging $39 an hour worked 37 hours and submitted an invoice for $1513. You'd probably not scrutinize it – since it is close to, and less than, $1600 (i.e., 40 × 40). It may be adaptive to assume that

[14] Of course, subjects can *learn* to solve the problem, with the benefit of instruction (Boissin, Caparos, Raoelison, & De Neys, 2021; Hoover & Healy, 2017) or exposure to multiple variants in succession (Raoelison et al., 2021; Raoelison & De Neys, 2019).

[15] Though not shown in Figure 2, we had an eleventh and twelfth condition in which price differences were $34 and $54 (total N= 2050). Contradicting the notion of "desirable difficulties" (Bjork, 1994; Alter et al., 2007; Mastrogiorgio and Petracca, 2014), performance was worse (26% and 24% correct) than the nearest conditions with round numbers. For further discussion of disfluency effects on solution rates, see Meyer et al. (2015) or Lawson, Larrick, and Soll (2022).

[16] We are not the first to observe effects of the numbers specified on the problem's solution rate, and these other results are also consistent with our "approximate checker" hypothesis. Frederick (2005) found improved performance when the items summed to 37 cents and differed by 13 cents. Baron et al. (2015) found dramatically improved performance when the two objects summed to $5.50 and differed by $1.00. Silva (2005) found that just 8% of his sample could solve the standard problem, whereas 93% could do so when the two objects summed to 3 cents and differed by 1 cent. However, unlike the aforementioned manipulations, Silva's result is probably *not* explained by our approximate checking hypothesis, since, with these values, subtracting the smaller number from the larger one actually yields the price of both objects (2 & 1) and the problem stipulates that the ball is the cheaper of the two.

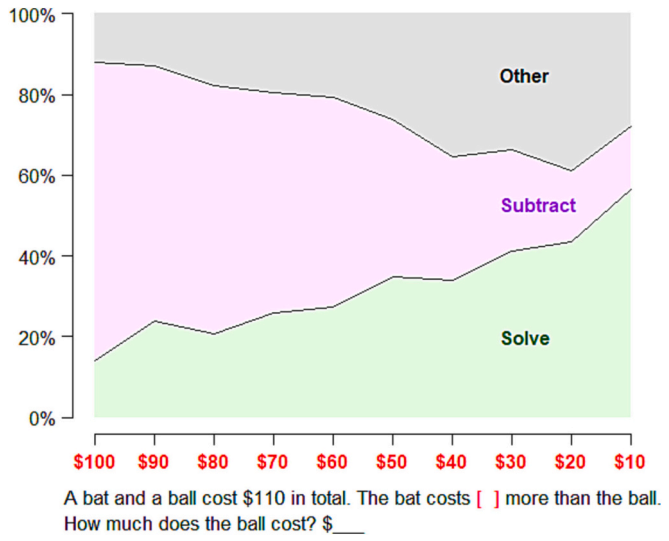


Fig. 2. Effect of price difference on percent of respondents who: Solve the problem, Subtract the price difference from $110, or give some Other incorrect answer.

plausible answers are correct, and it follows that people will have greater difficulty solving problems when the intuitive error is *approximately* correct.

The results of the prior study dovetail with other dual-process research on conflict detection, which finds, among other things, that base-rates are more likely to be incorporated into judgments if they are sufficiently extreme. For instance, if told that Bill likes carpentry, judgments of the likelihood that he is an engineer (vs. a lawyer) often neglect whether the relevant population is mostly engineers or mostly lawyers (Kahneman & Tversky, 1973). However, if the base rates are made sufficiently disparate (if only 5 of 1000 people in the sample are engineers) respondents *do* consider them; they notice (and must then resolve) the conflict between the disparate base rates and their assumption that engineers are more inclined towards carpentry (De Neys & Glumicic, 2008; Pennycook, Fugelsang, & Koehler, 2015).

In the context of the bat and ball problem, the cued operation (subtracting the smaller number from the larger one) only creates a *conflict* if respondents *notice* the constraints "hidden" in the problem stem. Respondents' superior performance when the cued operation yields an answer that more radically violates the problem constraints suggests that those constraints were never *completely* hidden (otherwise, the *degree* of violation wouldn't have *any* effect). Accordingly, although most subjects in our within-subject Hint experiments who initially say 10 cents either maintain that response ($n = 877$) or revise to 5 cents ($n = 731$), the sizable remainder who do neither ($n = 490$) are *much* more likely to adjust down to nine cents ($n = 113$) than up to eleven cents ($n = 14$). We interpret this as further evidence that they maintain *some* awareness of the constraint that they are still largely neglecting: that the two prices need to differ by $1.00 – and, correspondingly, that the correct answer must be *less than* 10. (See also, Bago, Raoelison, & De Neys, 2019).

Our approximate checker hypothesis suggests that even without being pressed to revise their initial response, a 9-cent ball will *feel* more correct than an 11-cent ball, because a 9-cent ball (and $1.01 bat) violate the $1.00 difference requirement less than an 11-cent ball (and 99 cent bat). We tested this conjecture in a study involving 1909 GCS respondents who were randomly assigned to one of nine conditions. In each, the standard bat and ball problem was presented along with two response options: the correct answer (5 cents) and an alternative value, X, which varied from 6 cents to 14 cents.[17]

Unsurprisingly, performance was much lower in the condition where X was the tempting lure (10) than in the other eight, more curious, conditions in which many respondents were forced to choose between two unintuitive options. But Fig. 3 further reveals that respondents perform worse if the alternative option is closer to 10, and worse for the four conditions with values *below* 10 (67%) than for the four with values *above* 10 (76%). Both of these additional results appear consistent with some version of our approximate checker hypothesis, though they remain distinct, as the first suggests scrutiny being withheld from responses that more closely resemble the dominant intuition (10), and the second suggests scrutiny being withheld from responses that more closely satisfy a requirement stipulated in the problem stem (that the bat and ball prices ought to differ by $1.00).

Of course, positing that the quality of an intuition is intuitively appraised is awkward, since it suggests that the intuitive "system" is checking itself. Yet the foregoing data do seem to argue against a fully deliberate checking process (in which violations of any degree would be equally wrong). Further, they appear consistent with the finding by Johnson et al. (2016) that, even under mnemonic load, respondents are

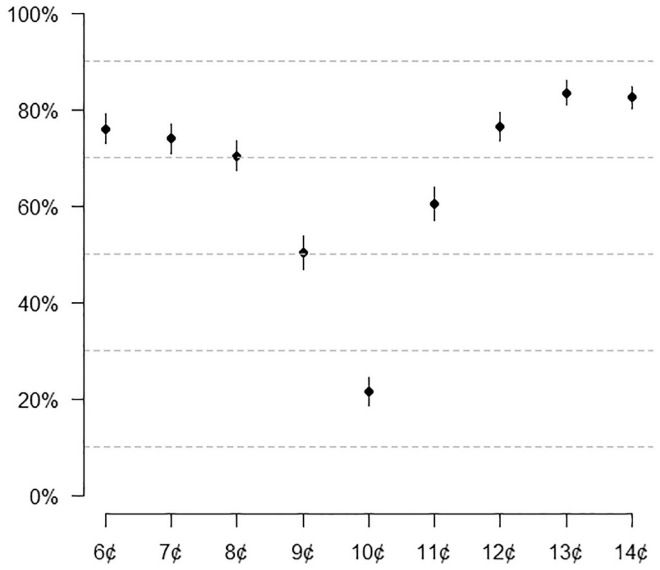


Fig. 3. % Choosing 5 cents over decoy for nine different decoys
Decoy varies between-subjects, forming nine binary-choice bat and ball conditions. Error bars indicate standard errors of the mean.

[17] This study was inspired by an episode of the game show "Who Wants to be a Millionaire?" which aired on November 10th, 2014. In that episode, the contestant (Erin LaVoie) correctly answered her first round question (regarding the meaning of the word "contusion") and then received the following question in round 2: "Try this tricky math question that stumps many Ivy Leaguers: A bat and ball cast $1.10. The bat costs $1 more than the ball. How much does the ball cost?" **(A)** $0.30 **(B)** $0.20 **(C)** $0.15 **(D)** $0.05. Curiously, this set of response options omitted (or excluded) the typical error ($0.10). With no obvious answer present, Erin first used her "Plus One" lifeline (in which a friend in the audience joins her to offer assistance), but after receiving insufficient help, then decided to use her "Jump the Question" lifeline to skip the question, foregoing her payoff from answering correctly ($5000), but removing her risk of answering incorrectly (which ends the game).

less confident in their 10-cent responses to the standard problem than in their 10-cent responses to the lite variant.[18] (See Appendix K for further data and discussion of the approximate checker hypothesis.)

4. General discussion

The title of a recent best-seller, *Thinking: Fast and Slow,* reflects the view that different types of cognitive processes can and should be distinguished (Kahneman, 2011). Others question the value of such an endeavor (Keren, 2013; Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011; Melnikoff & Bargh, 2018). The debate surrounding dual system theories is energized partly by differences in use of the term *theory* – whose meaning can range from a preliminary notion to a precisely stated and falsifiable hypothesis. Many dual system theorists regard their "theories" as provisional frameworks that help characterize and organize distinctions they find important, whereas critics may demand the sort of precision that would permit a decisive test of whether cognition has one system or two.[19]

Deciding how many mental 'systems' to enumerate depends on the sorts of distinctions one wishes to emphasize. If contrasted with, say, the digestive system, nearly everyone would attribute all thoughts to a single cognitive "system." But finer distinctions may also be useful, as Shweder (1977) so eloquently expresses:

> A useful distinction in the study of human thought is between intuitive and non-intuitive concepts. Concepts can be arranged along a continuum having to do with the relative ease with which they can be attained and in the kinds of learning inputs and environmental orchestration that are required for acquisition and application to occur… More intuitive concepts are acquired even under highly degraded learning conditions…[and] seem to be available without conscious effort or reflection…these concepts seem to be merely "released" by experience… In contrast, nonintuitive concepts require special learning conditions for their acquisition (e.g. massive instructional input, an orderly and explicit organization of learning trials, high motivation, etc.)

In our view, advocacy of dual process theories is typically nothing more (or less) than an endorsement of the possible value of distinguishing the types of thought a stimulus might generate or require; a desire to characterize mental operations in term of their speed, the amount of attention they demand or consume, their accessibility to introspection, and their difficulty of acquisition. Of course, not every distinction warrants the application of different labels: subjects would solve 7×12 faster than they would solve 18×27, but a book entitled *Thinking: Fast and Slow* would not be very compelling if these were the only sorts of data cited in support of the eponymous distinction.

Though the bat and ball problem has often been upheld as emblematic of the dual system framework, it may not be as canonical as its frequent citation suggests. First, the posited heuristic operation – subtraction – is, itself, a "rule-based manipulation of symbols," which is ordinarily ascribed to "System 2" (Sloman, 1996, p. 4). Second, this operation is sensitive to mnemonic load (DeStefano & LeFevre, 2004), which is sometimes taken as *the* defining feature of "Type 2" processes (Evans & Stanovich, 2013). Third, although intuitions are often defined by their lack of introspective access, those who miss this question know exactly how they arrived at their answer ($1.10 minus $1.00 equals $0.10); what they have trouble understanding is why that operation is inappropriate *here*.[20] Bago and De Neys (2019) further suggest that the 5-cent solution may *not* require deliberation (though we are skeptical of this claim, as discussed in Appendix L).

Further complications with dual system models arise when intuitions override reflection. Consider a version of the classic Monty Hall problem offered by Margolis (1987):

> Two Queens and a King are taken from a deck of playing cards, placed face down and shuffled. If you select the King, you win a prize. You first point to a card. The dealer then checks the two remaining cards and turns over a Queen. You may either keep the card you first pointed to or select the other card that remains face down. Is there any advantage to switching?

Since there are two ways to lose but just one way to win, you will be pointing to a losing card two out of three times. In both of those two cases, the remaining card that the dealer has not turned over will be the King. Thus, *if you switch*, you'll double your chance of winning: from 1 in 3 to 2 in 3.

Though few can offer any sensible rebuttal to this logic, it does not typically unseat the dominant intuition. Many conclude that they are missing something – that there's been some sleight of hand (Margolis, 1987). This turns the usual dual-process story on its head. If people remain incredulous following exposure to logic they cannot rebut, System 1 is effectively checking and overriding System 2. We find something similar with the bat and ball problem, as respondents seem to maintain a belief in a 10-cent ball (and $1.00 bat), despite having had their attention directed to the requirement that those two prices must *differ* by $1.00.

Intuitive answers may be even more influential for problems lacking any promise of an algorithmic solution. Consider our "minor injuries" problem, below:

> The Department of Transportation is deciding between two different roadway designs. These are associated with different types of auto accidents, and, consequently, with different rates of serious injuries and minor injuries. Please enter the number of minor injuries that would make the two designs equivalent, all things considered.

	Serious injuries	**Minor injuries**
Design A:	2000	16
Design B:	1000	____

By design, this problem has multiple compelling lures (8, 32, & 1016). Though all are absurd upon reflection (as all imply that minor injuries are as bad or worse than serious ones), their presence, coupled with the absence of any obvious alternative solution strategy, makes this problem difficult. Indeed, Meyer et al. (2023) find that fewer than one in fifty "solve" it (respond with a number above 1016). Moreover, this

[18] Our proposal that checking may actually be as intuitive as production of the intuition itself concords with research summarized by De Neys and Bonnefon (2013), who find that that those who succumb to the intuitive errors on classic heuristics and biases problems (such as Linda) are less confident than those who perform similar operations to correctly solve easier variants (De Neys, Rossi, & Houdé, 2013), take longer to respond (De Neys & Glumicic, 2008), show greater autonomic activation (De Neys, Moyens, & Vansteenwegen, 2010), and greater activation in brain regions supposed to mediate conflict detection (De Neys, Vartanian, & Goel, 2008; Simon, Lubin, Houdé, & De Neys, 2015).

[19] When a skeptic challenged J.B.S. Haldane to explain how evolutionary theory could be falsified, he famously shot back "*Fossilized rabbits in the Precambrian.*" (None have yet been found.) It is difficult to imagine a dual-system theorist producing a comparably pithy answer to a similar challenge.

[20] Further, the modest arithmetic abilities required to generate the intuition (i.e., subtraction) presumably correlate positively with the more demanding abilities required to solve the problem, thereby failing the "stochastic independence" criterion proposed by Tulving (1985), which is honored by those who propose distinct systems in the context of vision (Weiskrantz, 2009; Weiskrantz et al., 1974) or memory (Tulving, Schacter, & Stark, 1982).

problem may more closely resemble those we commonly confront, which *lack* established algorithms that might be used to check or override an intuition. Thus, the difficulty of inducing respondents to reflect on their 10-cent ball answer may understate the difficulty of inducing reflection more generally.

Although we've focused here on trying to understand why people typically miss the bat and ball problem rather than why their failure or success predicts other traits, the two issues are obviously related. We've proposed that performance on this item (and other items intended to measure cognitive reflection) is determined by the ability to detect and reject the erroneous intuition *and* by the ability to solve the problem once the error is detected. To help distinguish these two abilities, in some of our studies, participants first responded and were then told that the answer is not 10. As noted earlier, this "hinted" procedure serves to partition respondents into three groups: the *reflective* (who reject the common intuitive error and solve the problem on the first try), the *careless* (who answer 10, but revise to 5 when told they are wrong), and the *hopeless* (who are unable or unwilling to compute the correct response, even after being told that 10 is incorrect).

Expressed or implied claims that items intended to measure cognitive reflection have surplus predictive validity over other "regular" math problems suggest that the ability to suppress an activated intuition is an important cognitive skill distinct from numeracy or mathematical ability (Frederick, 2005). While some have affirmed this claim (Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2016; Shenhav, Rand, & Greene, 2012; Toplak, West, & Stanovich, 2011) others have disputed it (Attali & Bar-Hillel, 2020; Otero, Salgado, & Moscoso, 2022). Our analysis of the hinted procedure does suggest that these are dissociable skills. Specifically, as shown in Table 3, the *careless* perform nearly as well as the *reflective* on a subset of Raven's Matrices,[21] but nearly as poorly as the *hopeless* on the "Linda" problem.[22] This suggests that the bat and ball problem predicts Raven's scores because it requires mathematical ability (which the *careless* and *reflective* both possess), but predicts performance on the Linda problem, because it also requires the ability to suppress an activated intuition (which the *careless* and *hopeless* both lack).

More generally, we predict that items intended to measure cognitive reflection will be superior predictors for tasks which demand – and benefit from – an ability to detect and suppress a dominant intuition (such as Linda and other counterintuitive problems), but will function like "regular" math items for most other tasks whose items generally fail to induce a dominant intuition that must be suppressed (most numeracy tests, GRE math, and, perhaps, Raven's Matrices). Since providing the hint nullifies the importance of detecting the intuitive error, it goes some way to transform the CRT into a "regular" math test and we'd then expect it to function more like those tests. This general claim is supported by Table 3, by Appendix M, and by subsequent work (Meyer et al., 2023).[23]

If the bat and ball problem does measure anything distinct from general mental ability – such as willingness to reason carefully – we assume that it does so by permitting those disinclined toward reflection a chance to exit early while still feeling successful. Many other problems share this feature, such as the "XYZ" problem below.

$$\begin{aligned} x + y + z &= 1 \\ x^2 + y^2 + z^2 &= 2 \\ x^3 + y^3 + z^3 &= 3 \\ x^4 + y^4 + z^4 &= ? \end{aligned}$$

Since the universal intuition is obvious, we presume that nearly everyone who "solves" this item feels some sense of success. However, it offers little opportunity for the reflective to differentiate themselves, as negligibly few will have the motivation or ability to check whether this system of equations mutually entail the intuitive answer.[24]

Ideally, any item advanced as a measure of cognitive reflection should satisfy four criteria. It should (a) generate an intuition, which (b) requires suppression. Further, it should (c) contain cues to reject that intuition, and (d) allow those who do so to solve the problem.[25] The XYZ problem clearly satisfies (a) and (b), but clearly fails (c) and (d). By contrast, the bat and ball problem often satisfies all four criteria. It clearly satisfies (c), as it explicitly states that the two prices must differ by 100. For elite populations, it also satisfies (d). But elsewhere it does not, as many cannot solve the problem even when told that 10 is not the answer.

To the extent that problems satisfy these four criteria, we'd expect wrong answers to be generated more quickly than correct answers and highly concentrated at the putative intuition. Further, we'd expect those who miss such problems will judge them as easier than those who solve them (see Frederick, 2005; Mata et al., 2013). To help illustrate these criteria, consider three logically equivalent variants of a novel item that we call the "smokers" problem.[26]

Table 3
Raven's and Linda performance by Bat and Ball response $_{\text{Number of observations}}$

	Raven's score (out of 6)	Avoiding Conjunction Fallacy
Reflective (5 on first try)	3.5 $_{551}$	34% $_{439}$
Careless (10, but revised to 5)	3.2 $_{427}$	22% $_{317}$
Hopeless (10, and never got 5)	2.4 $_{881}$	19% $_{563}$

[21] We used items 2, 8, 14, 20, 26, and 34 from Raven's Advanced Progressive Matrices.

[22] Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. Which is more probable? Linda is a bank teller. OR Linda is a bank teller and is active in the feminist movement.

[23] We exclude from Table 3 the small minority (about 5%) who initially gave atypical answers for the ball's price (210, 105, 50, etc.). Though such responses are often negligibly rare, Pennycook, Cheyne, Koehler, and Fugelsang (2016) note that the CRT has greater predictive validity when it is scored in terms of number correct (5, 5, and 47) than when (reverse) scored as number of intuitive responses (10, 100, and 24). Expressed differently, they find it better to treat atypical answers as incorrect answers (failed to solve it) than as correct answers (managed to avoid intended trap). Some interpret this as evidence against the traditional dual-process interpretation of the CRT, which places emphasis on surmounting the intuition. We aren't fully persuaded by this critique, as these atypical answers could instead reflect submission to some unintended lure or to corruption of subsequent thinking by temporary consideration of the intended lure, even after it has been (partially) dismissed.

[24] The correct answer is 25/6.

[25] If an item possesses these characteristics, the presence of a correct answer is sufficient evidence that an intuition has been rejected (i.e., that a person is "reflective"), but it is not the *only* kind of evidence. For instance, suppose Adam said, "I first thought the ball cost $10, but then I realized that can't be right, because then the bat, itself, would cost $110. I left it blank because I just couldn't figure out what the right answer is, but now I can't stop thinking about it." He's clearly reflective (if not especially numerate). So is Beth who answered $105, which is clearly not a thoughtless error (as revealed by response times and common sense). Carl might also be considered reflective if he answered $10 but indicated very low confidence in his answer. Though all missed the problem, each displayed reflection: they all appeared to notice and care about facts that conflicted with their (likely) intuition. Thus, "requirement" (d) is more pragmatic than essential. To the extent an item satisfies criteria (a) and (b), a correct answer is good evidence that an intuition has been suppressed – but not the *only* sort of evidence.

[26] This problem was inspired by an example provided by Chris Chabris.

#1: If 3 in 30 men smoke and 1 in 30 women smoke, then 1 in __ people smoke. (48% correct)

#2: If 1 in 10 men smoke and 2 in 60 women smoke, then 1 in __ people smoke. (9% correct)

#3: If 1 in 10 men smoke and 1 in 30 women smoke, then 1 in __ people smoke. (7% correct)

Since these variants all require averaging the same two fractions (1/10 and 1/30), they are *logically* equivalent. But they are *psychologically* distinct, as revealed by marked differences in responses, response times, and judged difficulty. Variant #1 fails as a measure of reflection because the intuitive operation (averaging numerators) "happens" to yield the correct answer. Variant #2 fails because it evokes no intuition; no simple operation promises to yield a solution. Variant #3 is more promising because it (a) suggests an intuitive operation (averaging denominators) which (b) yields an erroneous solution, such that reflection *will be* required.[27]

Fig. 4 plots the responses to these three variants by their response times and judged difficulty. Variant #1 is answered quickly and judged to be easy. Variant #2 is answered slowly and judged to be difficult. Only variant #3 resembles the bat and ball problem, as wrong answers are highly concentrated (almost all 20 or 40) and emitted much more quickly than correct answers. Moreover, those who miss this variant regard it as easier than those who solve it (note that the red 20 and red 40 in the bottom left are well below the red 15 in the top right).

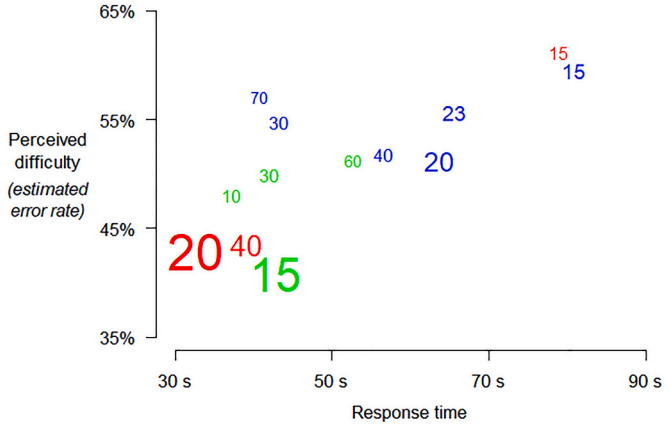


Fig. 4. Responses to smokers variant #1, #2, and #3
These data are from 1342 Mturkers randomly assigned to receive one of the three variants. For each condition, we graphed all responses given by at least 5% of respondents, with font size proportional to response share.

5. Concluding remarks

When we began studying the bat and ball problem, we assumed respondents missed it because they didn't bother to check. Accordingly, we assumed that they'd be able to solve it if we directed their attention to the features of the problem that differentiate it from the problem we thought they were unwittingly solving instead (bat and ball "lite") or to the constraint the typical answer violates (that the prices differ by 100).

We discovered instead that many respondents maintain the erroneous response in the face of facts that plainly falsify it, even after their attention has been directed to those facts. Although subjects' apparent sensitivity to the *size* of the heuristic error merits further research, the remarkable durability of that error paints a more pessimistic picture of human reasoning than we were initially inclined to accept; those whose thoughts most require additional deliberation benefit little from whatever additional deliberation can be induced.

CRediT authorship contribution statement

Andrew Meyer: Conceptualization, Writing – original draft, Writing – review & editing, Visualization, Data curation, Investigation, Formal analysis. **Shane Frederick:** Conceptualization, Writing – original draft, Writing – review & editing, Visualization, Supervision, Funding acquisition.

Data availability

Data will be made available on request.

Appendix A. The hijackers question

Were the 9/11 hijackers cowards? YES NO

Since it is tempting to apply a negative label to a negative stimulus, we assumed that the intuitive answer is "Yes" and the reflective answer is "No." Our data supported this conjecture. Among 497 GCS respondents, the minority who answered "No" (24%) took two seconds longer to respond (7.0 vs. 5.0 s), and they solved the bat and ball problem at higher rates (30% vs. 18%). A follow up study on MTurk ($N = 1678$) yielded comparable results, as the minority who answered "No" (34%) took longer to respond (8.4 vs. 7.0 s), and were more likely to solve the bat and ball problem (58% vs. 39%).

[27] However, the "Smokers" problem likely fails criterion (c), because it isn't obvious what the modal intuitive response might be checked *against*. It likely also fails criterion (d), because even if a respondent *did* experience conflict ("Wait a second, it would be 1/20 if ***no*** women smoked…"), they may still be unable to solve the problem.

Appendix B. Response latencies

Table B
Response times for the most common responses.

Response	response time (s)	% who responded in less than			
		5 s	10 s	20 s	40 s
\$0.10 $_{n=1315}$	28	0	2	34	78
\$0.05 $_{n=288}$	57	0	1	8	37
\$10.00 $_{n=74}$	39	1	7	24	51
\$1.00 $_{n=62}$	23	2	21	44	81
\$2.10 $_{n=35}$	21	0	11	63	89
\$0.00 $_{n=33}$	33	0	0	36	52
\$1.05 $_{n=23}$	45	0	0	0	43
Other $_{n=170}$	24	3	21	45	72

From a sample of 2000 U.S. web-browsers collected by GCS.

Appendix C. Demographics

Table C
Age and gender for studies in the empirical sections of the main text

Section	Study	% Male	Mean Age
1	Both Bat and Ball	61	28
1	Recall	58	37
1	Bat	57	42
1	Difference	55	45
2	Not sure	59	45
2	Warnings	55	42
2	Hints	54	35
2	Subtle Confirmation	52	44
2	Confirmation	40	46
2	Suggesting 5	56	45
3	Small difference	50	42
3	Crazy alternative	60	45

Appendix D. Remembering the problem

As predicted by the attribute substitution hypothesis, respondents who answer 10 cents are more likely to mis-recall the problem as its "lite" variant. But contrary to the attribute substitution hypothesis, most 10-cent respondents *do* recall the problem correctly. This was true whether we used the free *recall* task described in the main text or a *recognition* task in which respondents were presented with the bat and ball problem and its lite variant and indicated which of the two problems they had previously answered (see Tables D1 and D2). These results comport with those reported by Hoover and Healy (2019), by Mata, Schubert, and Ferreira (2014), and with an eye tracking study, which found that solution rates were higher among those who spent more time looking at the differentiating detail (Mata, Ferreira, Voss, & Kollei, 2017).

Table D1
Recall

Response to Bat and Ball problem	% recalling Correctly	% mis-recalling as lite variant	% making other mnemonic errors
5 cents $_{n=158}$	94	0	6
10 cents $_{n=397}$	61	23	16
other $_{n=60}$	43	7	50

Table D2
Recognition

Response to Bat and Ball problem	% recognizing Correctly	% mis-identifying as lite variant	% making other mnemonic errors
5 cents $_{n=164}$	100	0	-
10 cents $_{n=225}$	76	24	-
other $_{n=27}$	81	19	-

Appendix E. Emphasizing the difference between bat and ball and the posited substitute

The main text reports results from a study in which we emphasized the difference between the bat and ball problem and its hypothesized substitute by bolding the critical words (see the first condition below). Two other manipulations are not reported in the main text. One juxtaposed the problem against its hypothesized substitute. The other rephrased the standard problem in a way we thought might reduce the likelihood of the hypothesized substitution.

BOLD
A bat and a ball cost $1.10 in total. The bat costs $1.00 **more than the ball**.
How much does the ball cost? ____

CONTRAST WITH LITE
A bat and a ball cost $1.10 in total. The bat costs $1.00.
How much does the ball cost? ____

A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball.
How much does the ball cost? ____

REPHRASED
A bat and a ball cost $1.10 in total. Their prices differ by $1.00. The ball is cheaper.
How much does the ball cost? $__

Table E reports the (null) results of these manipulations across four studies totaling 5479 participants. The null effect of the "Contrast with Lite" manipulation contradicts the results of Hoover and Healy (2021). The null effect of the bolding and rephrased manipulations converges with a study run by Bourgeois-Gironde and Van Der Henst (2009).

Table E
Effects of attempts to impede the substitution

	A. eLab $N = 208$	B. High School[a] $N = 228$	C. Google[b] $N = 2956$	D. Google $N = 2087$
Control	32 $_{63}$	72 $_{22}$	20 $_{-}$	10 $_{77}$
Bold	24 $_{53}$	–	21 $_{-}$	–
Contrast with lite	36 $_{52}$	71 $_{23}$	–	–
Rephrased	–	–	–	11 $_{71}$

Main script indicates percent correct. Subscript indicates percent 10 cents.
[a] We thank Elizabeth Zhou for collecting these data and for obtaining permission to run the study.
[b] In this study, rather than generating a response, participants responded by choosing between 5 and 10 cents.

Mata et al. (2013) and Mata (2020) conducted within-subject versions of the bolding manipulation. Their participants answered the bat and ball problem, submitted their answers, and then encountered the problem again with the critical text now underlined. Both papers report small performance increases in each of four experiments. We attempted a near exact replication, assigning 106 eLab participants to the same within-subject experiment. But we found no change in performance. 26% of participants got the problem right, both initially and after the second chance with the critical text underlined.

A follow up study on *GCS* ($N = 252$) helps explain why these manipulations are ineffective: most respondents fail to appreciate the significance of the words "more than", even when the standard problem is explicitly contrasted with its simpler "lite" variant.

#1) A bat and a ball cost $1.10 in total. The bat costs $1.00.
#2) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball.

Which is true?

A) The ball costs different prices in #1 and #2. **(43%)**
B) The ball costs the same price in #1 and #2. **(57%)**

Appendix F. Using bat price item to diagnose three levels of reasoning

On page 3 of the main text, we showed that if asked for the price of the *bat*, respondents make two types of errors: some treat the second number as the bat's cost (a *substitution* error), whereas others simply subtract the smaller number from the larger one (a *subtraction* error). We found similar results in a study conducted on GCS ($N = 1019$) using a variant with novel names and prices.

A clabor and a plonket cost $330 in total.
The clabor costs $300 more than the plonket.
How much does the [Clabor / Plonket] cost? $___

The problems were similarly difficult, regardless of which price was sought. However, if asked for the price of the plonket, the only common error was $30 (offered by 68% of respondents), but when asked about the price of the clabor, we again found two types of errors, with 32% answering $300 (a substitution error) and 32% answering $30 (the subtraction error). As with the study reported in the main text, the subtraction error appeared to be "more intuitive" as it was emitted significantly faster (23 vs. 40 s), suggesting (even) less reflection and shallower reasoning than the substitution error. A follow up study on MTurk ($N = 1713$) supports this interpretation, as those who "solve" the item by subtracting the smaller number from the larger one show the weakest reasoning skills on a battery of other items intended to measure degree of reflection (see materials below & Table F).

Bat Price item:
A bat and a ball cost $330 in total. The bat costs $300 more than the ball.
How much does the **bat** cost? $___

Battery of other reasoning items:

- Were the 9/11 hijackers cowards? YES NO
- Could Adolf Hitler be considered successful in some respects? YES NO
- What percentage of seven-letter English words have the following forms?
  _ _ _ _ _ N __%
  _ _ _ _ I N G __%
- On what day does "early July" become "mid-July"? July ___
- If it takes 6 machines 6 minutes to make 6 widgets, how long would it take 60 machines to make 60 widgets? ___ minutes
- In a lake, there is a patch of lilypads. Every day, the patch doubles in size.
  If it takes 44 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? ___ days
- A basket contains apples that are either red or green. Sixty of the apples are red. Forty percent of the apples are red.
  How many green apples are in the basket? ___
- A bat and a ball cost $330 in total. The bat costs $300 more than the ball. How much does the **ball** cost? $___
- A bat and a ball cost $330 in total. The bat costs $300. What is the price difference between the bat and the ball? $___
- If it takes 5 elves 6 minutes to wrap 5 presents, how long would it take 60 elves to wrap 60 presents? ___ minutes
- Which comes closer to your view?
  _Humans and other living things have existed in their present form since the beginning of time.
  _Humans and other living things have evolved over time.

Table F
Relation between bat price answer and responses to other questions[a]

	bat price answer		
Percent who believe that…	$30 $_{n=461}$	$300 $_{n=339}$	$315 $_{n=508}$
9/11 hijackers are NOT cowards	25	24	47
Hitler was successful	62	64	75
_N_ is more common than ING	21	28	48
Mid July begins on the 10th–14th	24	31	61
Ball costs $15	2	2	95
Machines take 6 min	38	43	85
Lilypads take 43 days	35	42	89
There are 90 green apples	19	30	66
Bat and ball differ by $270	17	33	70
Elves take 6 min	26	36	71
We evolved	69	75	87

[a] Excluded here are 405 respondents who gave other answers, of which 275 were $15 (the correct price for the *ball*). This error was much more common in our Mturk sample than in our GCS sample, perhaps because MTurkers have been repeatedly exposed to the standard problem (which asks for the price of the ball).

Appendix G. Analyses of respondents' confidence in their answers

In some of studies we conducted on MTurk and eLab, we elicited respondents' confidence in their answers. Most (73%) who said 10 cents ($N = 773$) were *maximally* confident, selecting 100% from an 11-point scale ranging from 0% to 100%. Further, 16 of the 22 Yale undergraduates who made the common error were sufficiently confident to bet on it, preferring to receive $2 for a correct response than $1 for sure.[28] A subsequent multiple-choice study conducted on GCS ($N = 602$), shown below, further revealed that the intuitive error is held with considerable confidence even in the presence of the correct response, and it was unaffected by a warning that the problem was more difficult than it appears.[29]

CONTROL CONDITION		***WARNING CONDITION***	
		Be careful! Many people miss this problem.	
A bat and a ball cost $110 in total. The bat costs $100 more than the ball.		A bat and a ball cost $110 in total. The bat costs $100 more than the ball.	
How much does the ball cost?		How much does the ball cost?	
○ $5	**16%**	○ $5	**19%**
○ $10	**63%**	○ $10	**59%**
○ Not sure	**21%**	○ Not sure	**22%**

Appendix H. Evaluating pairs of prices

As reported in the main text (and reproduced below), respondents are much more inclined to affirm that a $100 bat is $100 more than a $10 ball if the $110 sum is mentioned in the problem stem.[30]

MINUEND ABSENT		
A bat costs $100 more than a ball.		
	18%	
If you said the bat cost $100 and the ball cost $10, would your prices be correct?	YES	NO
MINUEND PRESENT		
A bat and a ball cost $110 in total. The bat costs $100 more than the ball.		
	53%	
If you said the bat cost $100 and the ball cost $10, would your prices be correct?	YES	NO

We assume that the presence of $110 invites respondents to subtract $100 from it, yielding the $10 ball mentioned in the second statement, which increases their confidence in that statement to a degree that impairs their ability to notice that a $100 bat and $10 ball don't satisfy the other constraint (differing by $100). A follow up study ($N = 1003$; GCS) supports this account, as mentioning the $110 total no longer induces errors if the ball price stated in the second sentence does not match the difference between $110 and $100.

MINUEND ABSENT		
A bat costs $100 more than a ball.		
	11%	
If you said the bat cost $90 and the ball cost $20, would your prices be correct?	YES	NO
MINUEND PRESENT		
A bat and a ball cost $110 in total. The bat costs $100 more than the ball.		
	15%	
If you said the bat cost $90 and the ball cost $20, would your prices be correct?	YES	NO

A second follow up study shown below ($N = 2000$; GCS) reveals that explicit provision of the *difference* constraint also increases errors, but that the

[28] The intuitive error was held with much greater confidence than atypical errors (such as $2.10). Among the 100 respondents who committed atypical errors "only" 46% were maximally confident, and just 1 of the 3 Yale students who committed an atypical error was willing to bet on her response. Confidence in the intuitive error was however, not *quite* as high as confidence in the correct answer. Among the 433 respondents who said 5 cents, 83% were maximally confident. Moreover nearly all (106) of the 109 Yale undergraduates who responded correctly were willing to bet on their answer. (See also De Neys et al., 2013.)

[29] Respondents were, of course, randomly assigned to one of these conditions. Data from the control condition were nearly identical to an earlier study conducted on GCS ($N = 808$), which involved just that condition.

[30] A futher condition ($N = 500$; GCS) showed that the error rate was just as high (57%) when the order of the two constraints is reversed.

effect is much smaller. In other words, respondents are much less likely to endorse a pair of prices that violate the *total* constraint than the *difference* constraint (28% vs. 53%).[31]

SUBTRAHEND ABSENT
A bat and a ball cost $110 in total.

	15%	
If you said the bat cost $110 and the ball cost $10, would your prices be correct?	YES	NO

SUBTRAHEND PRESENT
A bat and a ball cost $110 in total. The bat costs $100 more than the ball.

	28%	
If you said the bat cost $110 and the ball cost $10, would your prices be correct?	YES	NO

A final study ($N = 2005$; GCS) showed that presentation of the $110 sum not only blinds respondents to the fact that $100 and $10 do *not* differ by $100, but somewhat impairs their ability to correctly conclude that $105 and $5 *do* differ by $100.

MINUEND ABSENT
A bat costs $100 more than a ball.

	71%	
If you said the bat cost $105 and the ball cost $5, would your prices be correct?	YES	NO

MINUEND PRESENT
A bat and a ball cost $110 in total. The bat costs $100 more than the ball.

	56%	
If you said the bat cost $105 and the ball cost $5, would your prices be correct?	YES	NO

Appendix I. Does 100 minus 10 equal 100?

In the main text, we reported the first (*Provided Prices*) and last (*Generated Prices*) conditions of a four-condition experiment summarized below (total $N = 247$). Results from the other two conditions are shown below. The *Generated Lite* condition is an additional control. It shows that merely generating the answers $0.10 and $1.00 is insufficient to yield the very high error rates. The version of the *Generated Prices* condition that maintains the *provided prices question* format shows that the effect reported in the main text is preserved after eliminating differences in wording between conditions.

PROVIDED PRICES
A bat costs $1.00 and a ball costs $0.10.

	6%	
With those prices, does the bat cost $1.00 more than the ball?	YES	NO

GENERATED PRICES LITE
A bat and a ball cost $1.10 in total. The bat costs $1.00.
How much does the ball cost? $ 0.10
How much does the bat cost? $ 1.00

	30%	
With those prices, does the bat cost $1.00 more than the ball?	YES	NO

GENERATED PRICES (provided prices question format)
A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball.
How much does the ball cost? $ 0.10
How much does the bat cost? $ 1.00

	67%	
With those prices, does the bat cost $1.00 more than the ball?	YES	NO

GENERATED PRICES
A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball.
How much does the ball cost? $ 0.10
How much does the bat cost? $ 1.00

	76%	
Is your "bat" answer $1.00 more than your "ball" answer?	YES	NO

We conducted the same experiment on MTurk (total $N = 176$), but this time first asked participants whether the prices summed to $1.10 before asking them whether they differed by $1.00. That substantially increased participants' erroneous endorsement of the $1.00 difference, but the large gap between the *Provided* and *Generated* conditions remains, replicating the result discussed above (and emphasized in the main text).

[31] This converges with results reported on page 2 in the main text, where we found that respondents asked to produce a bat price and a ball price were vastly more likely to violate the difference constraint than the total constraint.

PROVIDED PRICES
A bat costs $1.00 and a ball costs $0.10.
With those prices, do the bat and ball cost $1.10 in total? **100% "YES"**
With those prices, does the bat cost $1.00 more than the ball? **41% "YES"**

GENERATED PRICES LITE
A bat and a ball cost $1.10 in total. The bat costs $1.00.
How much does the ball cost? $ 0.10
How much does the bat cost? $ 1.00
With those prices, do the bat and ball cost $1.10 in total? **100% "YES"**
With those prices, does the bat cost $1.00 more than the ball? **49% "YES"**

***GENERATED PRICES* (provided prices question format)**
A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball.
How much does the ball cost? $ 0.10
How much does the bat cost? $ 1.00
With those prices, do the bat and ball cost $1.10 in total? **90% "YES"**
With those prices, does the bat cost $1.00 more than the ball? **85% "YES"**

GENERATED PRICES
A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball.
How much does the ball cost? $ 0.10
How much does the bat cost? $ 1.00
Do your two answers sum to $1.10? **100% "YES"**
Is your "bat" answer $1.00 more than your "ball" answer? **86% "YES"**

Appendix J. Suggestion experiments

We assigned 8026 GCS respondents to either a control condition or to one of five "suggestion" conditions (the final two of which were discussed in the main text). As shown in Table J, requests to "consider" an incorrect value had little effect, whereas requests to consider the correct value had a moderate effect. Unsurprisingly, explicitly telling participants that the answer was 5 had a large effect, but even then, a substantial minority insisted that the answer was 10.

Table J
Effects of suggestions

Problem text:	% "5"	% "10"
A bat and a ball cost $110 in total. The bat costs $100 more than the ball. How much does the ball cost?	13	73
A bat and a ball cost $110 in total. The bat costs $100 more than the ball. How much does the ball cost? **Before responding, consider whether the answer could be $33.**	13	63
A bat and a ball cost $110 in total. The bat costs $100 more than the ball. How much does the ball cost? **Before responding, consider whether the answer could be $10.**	19	66
A bat and a ball cost $110 in total. The bat costs $100 more than the ball. How much does the ball cost? **Before responding, consider whether the answer could be $5 or $10.**	26	64
A bat and a ball cost $110 in total. The bat costs $100 more than the ball. How much does the ball cost? **Before responding, consider whether the answer could be $5.**	31	55
A bat and a ball cost $110 in total. The bat costs $100 more than the ball. How much does the ball cost? **The answer is $5. Please enter the number 5 in the blank below.**	77	18

Appendix K. Further tests of the "approximate checker" hypothesis

The main text reported the results of a 10-condition study, which showed that if the stated price difference between the bat and ball decreases, respondents slow down and do better. Table K, below, shares the results from 11 similar experiments involving a total of 11,730 participants from five different populations. Though these other studies involved bats and balls whose prices were denominated in cents rather than dollars, they were otherwise nearly identical to the study reported in the main text. The problem always read "A bat and a ball cost $1.10 in total. The bat costs [X] more than the ball." The value of X was manipulated between conditions. Depending on the study, the specified price differences ranged from as high as $1.04 to as low as $0.10.

These other eleven experiments largely replicate the results from the study we reported in the main text. Moreover, they provide further support for our notion of an "approximate checker." First, when the heuristic operation (subtraction) creates larger errors, respondents are more likely to reject its result, causing them to spend more time on the problem, which elevates solution rates. Second, four of these eleven studies (A-D) show that when the stated price difference is smaller, those who cannot solve the problem begin feeling more uneasy about their intuitive answer. In studies A and B, participants reported confidence on a 5-point Likert scale ranging from "not at all confident" to "completely confident." In studies D and E, they reported it on an 11-point scale ranging from a 0% to 100% chance of being correct. Those who solved the problem were comparably confident whatever the specified price difference, but those who missed it were less confident when the price difference was smaller ($t = 4.2$).

Table K1
Effect of price difference between bat and ball

Price Difference:	A. eLab $N = 304$	B. Yale $N = 41$	C. Google $N = 3945$	D. mTurk $N = 560$	E. mTurk $N = 321$	F. Google $N = 2008$	G. Boston[a] $N = 534$	H. Google $N = 804$	I. Google $N = 204$	J. Google $N = 2004$	K. Google $N = 1005$
$1.04	–	–	–	–	–	–	–	–	10 $_{17}$	–	–
$1.02	–	–	–	–	–	–	–	–	–	–	13 $_{27}$
$1.00	29 $_{23}$	48 $_{14}$	19 $_{14}$	28 $_{32}$	32 $_{19}$	12 $_{16}$	32	–	12 $_{16}$	14 $_{23}$	18 $_{25}$
$0.88	45 $_{45}$	45 $_{30}$	26 $_{20}$	26 $_{56}$	–	–	–	–	–	–	–
$0.70	–	–	–	–	–	–	–	22 $_{26}$	–	–	–
$0.60	–	–	–	–	–	–	38	25 $_{28}$	–	–	–
$0.50	–	–	–	–	–	–	45	34 $_{30}$	–	–	–
$0.40	–	–	–	–	–	–	–	36 $_{39}$	–	–	–
$0.22	–	–	–	54 $_{96}$	–	–	–	–	–	–	–
$0.12	–	–	57 $_{18}$	–	–	–	–	–	–	–	–
$0.10	–	–	64 $_{18}$	63 $_{54}$	57 $_{34}$	46 $_{21}$	56	–	–	52 $_{30}$	–

Main script indicates percent correct. Subscript indicates seconds to respond.
[a] Response time data were omitted from this population because the experiment was done using paper and pencil.

To further analyze how the stated price difference affects response times, we also decomposed the data from the 10-condition price difference study reported in the main text, analyzing not just *average* response times across *all* participants, but also, response times for each of the three groups discussed there (see Fig. 2): those who *solve* the problem, those who *subtract* the smaller number from the larger one, and those who give some *other* answer. As discussed earlier, overall, decreasing the price difference increases response times ($t = -11.9$). For instance, respondents spend significantly longer on the problem if the prices differ by $10 than if they differ by $100 (34 vs. 26 s, $t = -6.9$). However, Figure K shows that if we restrict focus to those who *solve* the problem, this relation becomes an inverted-u. Correct responses require more time as the price differences fall from $100 to $60, but less time as they fall further to $10, whereas response time for errors is unaffected by the difference manipulation.

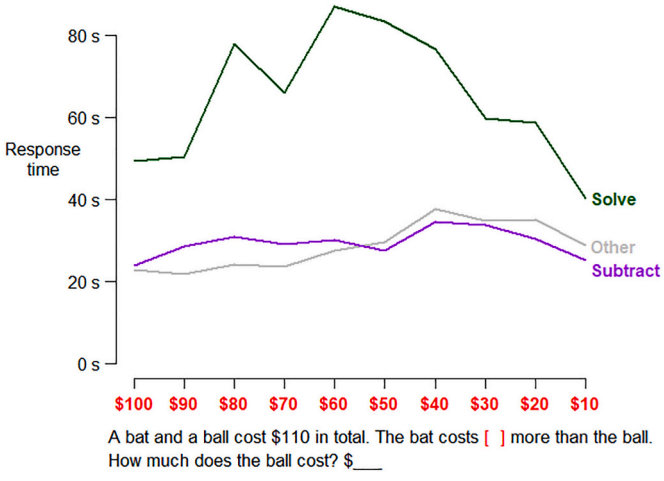


Fig. K. Effect of the stated price difference on response time

We've proposed the notion of an "approximate checker" which pardons small violations of the difference constraint, but not large ones. An alternative account is that smaller differences suggest a solution strategy, such as adjusting from half of the total until two values are found that satisfy both constraints. However, if the small difference variant primes some useful (and more general) solution strategy, we'd expect prior presentation of the small difference variant to increase performance on the standard problem, but we actually found the *reverse*. In a study involving 720 GCS

respondents, solution rates for the standard problem were not influenced by question order, but prior presentation of the standard problem lowered performance on the small difference variant.

STANDARD PROBLEM FIRST
A bat and a ball cost $110 in total. The bat costs $100 more than the ball.
How much does the ball cost? $___ **19% correct**

A bat and a ball cost $110 in total. But this time, the bat costs $10 more than the ball.
In this case, how much does the ball cost? $___ **40% correct**

SMALL DIFFERENCE FIRST
A bat and a ball cost $110 in total. The bat costs $10 more than the ball.
How much does the ball cost? $___ **60% correct**

A bat and a ball cost $110 in total. But this time, the bat costs $100 more than the ball.
In this case, how much does the ball cost? $___ **17% correct**

Another distinct alternate account of the price difference effect is suggested by Trémolière and De Neys (2014). If respondents expect bats to cost substantially more than balls, a smaller difference between item prices will cause the heuristic operation (subtracting the smaller number from the larger one) to yield prices that more strongly violate this expectation. However, we are skeptical of this interpretation, in part from the (non) results of a study we conducted on GCS in which 411 respondents were randomly assigned to one of the two conditions below.

AL & BOB
Al and Bob are 50 years old in total. Al is 20 years older than Bob.
How old is Bob? ___ **32% correct**

FATHER & SON
A father and a son are 50 years old in total. The father is 20 years older than the son.
How old is the son? ___ **36% correct**

Although the semantic content of the *Father & Son* condition would seem to invalidate the heuristic operation more forcefully (yielding a 20 year-old father with a 30 year-old son), solution rates are nearly unaffected. Accordingly, we doubt that the effects of the price difference on solution rates (see Fig. 2) reflect beliefs about the relative cost of bats and balls – and, further, we doubt that the heuristic error will typically be very sensitive to manipulations of the *semantic* content. (The prevalence of the subtraction error in the *bat price* variant discussed on page 3 provides further evidence of the neglect of semantic detail.)

Varying the *difference* between item prices is not the only way to manipulate the degree to which the intuitive operation violates the stipulated constraints; that can also be achieved by manipulating the specified *total*. For example, Baron, Scott, Fincher, and Metz (2015) report that just 38% solved the standard bat and ball problem (in which the two items sum to $1.10 and differ by $1.00), whereas 90% solved a "soup and salad" variant (in which the two items sum to $5.50 and differ by $1.00).

To further test the effect of manipulating the *total* price, we randomly assigned 1286 Prolific participants to one of seven conditions: a control condition (total of $110 and difference of $100), three conditions which hold the total price at $110 while reducing the difference, and three conditions which increase the specified total while holding the difference at $100. (Though we doubt it matters much, all conditions involved a "clabor" and a "plonket", to remove any semantic variance in how "realistic" the resulting prices were.)

Solution rates from these seven variants are shown in Table K2. The top row is the standard control condition (total = 110, difference = 100). The left side reports results of the variants that manipulated the *difference* (while holding the total constant). The right-side reports results of the conditions that manipulated the *total* (while holding the difference constant).

Regardless of whether manipulations involve the difference or the total, solution rates increase if differences are a smaller *proportion* of the total. While this provides further support for our approximate checker hypothesis, the effects are more modest than implied by the aforementioned "soup & salad" variant. Although we recognize the irony of suggesting this in the context of a 2nd table in a Kth appendix, perhaps "more research is needed" regarding the problem elements that do or do not matter.

Table K2
Further tests involving manipulation of problem elements

	110, 100: **35%**	
110, 60: **46%**		180, 100: **38%**
110, 50: **51%**		220, 100: **42%**
110, 10: **60%**		1100, 100: **65%**

Within each cell, the first number is the price total, the second number is the price difference, and the third number is the solution rate.

Appendix L. Reassessing the putative evidence for intuitive solutions

We assume that solving the bat and ball problem requires slow, effortful deliberation. Respondents who solve it take much longer than those who don't (see our Appendix B), and solution rates are markedly reduced by the imposition of time limits (Borghans et al., 2008) or mnemonic load (Johnson et al., 2016).

By contrast, Bago and De Neys (2019) propose that many can solve the problem intuitively. As part of their case for a "Smart System 1," they use a multi-trial, two-response paradigm in which respondents must first respond quickly under mnemonic load, but later get to respond again with no time pressure or load. They claim that most of those who ultimately solve the problem could do so intuitively (i.e., quickly, and despite cognitive load).

We are unpersuaded. First, their respondents don't just answer the standard bat and ball problem; they answer many slightly modified variants of the bat and ball problem interspersed among versions of "bat and ball lite." Many of the so called "intuitive" solutions are from these later trials; from variants of a problem respondents have already repeatedly encountered. We think it is important to distinguish *intuiting* an answer from quickly *applying* a solution strategy discovered during an earlier trial.

Our concern that repeated exposure exaggerates how many participants appear to *intuit* the solution is based, among other things, on our re-analysis of data in Raoelison and De Neys (2019), who used the two-response paradigm described above, and data from Raoelison, Keime, and De Neys (2021), who intermixed 4-s and 25-s trials in a single response paradigm. Fig. L pools the speeded responses from these two papers and plots solution rates by trial. On the first trial (red dot), only 1 in 30 respondents select the correct answer. The "intuitive" solution rate increases dramatically over the next forty or so trials, before falling slightly and leveling out close to 25% (which could be achieved by fatigued respondents randomly choosing one of the four response options).

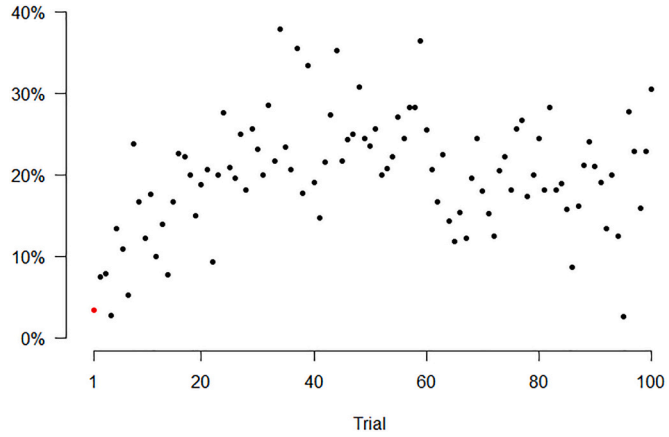


Fig. L. Solution rates for speeded responses from Raoelison and De Neys (2019) and Raoelison et al. (2021).

Second, most of these experiments, and the independent replication by Burič and Konrádová (2021), used a multiple-choice response format (in which the correct answer is presented alongside one, two, or three incorrect options). To illustrate our objections to this paradigm, suppose you put respondents under mnemonic load, enforced a six second time limit, used experimental instructions which alert respondents to the distinction between intuitive and deliberate responses, and then posed the "XYZ" problem, as below:

$$\begin{aligned} x + y + z &= 1 \\ x^2 + y^2 + z^2 &= 2 \\ x^3 + y^3 + z^3 &= 3 \\ x^4 + y^4 + z^4 &= ? \end{aligned}$$

Please indicate your very first, intuitive answer! 4 OR 25/6

Some may select 25/6 because the set up implies that there are two possible answers and 4 may seem suspiciously obvious. But we'd not conclude that *any* of these respondents were solving this set of equations – much less doing so within 6 seconds, under load.

When Bago and De Neys (2019) used an open-ended response format, they still found that 15 of the 50 respondents who eventually answered the first trial correctly could do so on the first of the two response opportunities (within about 6 seconds and despite the imposition of concurrent cognitive load). However, we'd regard that result as an anomaly.[32] First, our foregoing analyses revealed that only 1 in 30 respondents produced the correct answer on the first trial *even when it was included as one of the response options*. Second, when we presented the standard open-ended problem to

[32] We suspect that this result is a small sample fluke or reflects prior exposure to the problem. Like us, they excluded participants based on self-reported exposure to the problem. However, these self-reports should be interpreted cautiously. Meyer et al. (2018) found that many (1368 out of 4731) who claimed that they had never seen the problem before had, in fact, both seen and answered that identical problem *at least* once before (based on repeatedly appearing MTurk IDs).

a sample of 2000 American internet users who were not part of any regular participant pool that might have previously exposed them to it, 288 answered correctly. However, none did so within 6 seconds, and only eight did so within 12 seconds.[33] Further details are presented in Appendix B.

Furthermore, the putative "Smart System 1" is not correlated with cognitive ability in ways one might expect, or in ways that have been claimed. In Raoelison, Thompson, and De Neys (2020), cognitive ability was operationalized as participant's score on a 12-item Raven's APM and a 4-item "verbal CRT." Their claim that cognitive ability correlated positively with both initial (intuitive) accuracy and final (reflective) accuracy on the sequentially presented bat and ball variants holds only if later trials are included (which we find problematic for the aforementioned reasons).[34]

Table L presents correlations between bat and ball accuracy and cognitive ability for each trial within their second experiment.[35] On the *first* trial, only 2 of 54 participants initially selected the correct answer to the bat and ball problem. Both scored more than two standard deviations *below* the sample mean on their 16-item test of cognitive ability, and both switched to the wrong answer after deliberating, suggesting that many of the so-called intuitive solutions actually reflect an *inability* to even perform the intuitive calculation within the permitted time.[36]

Table L
Correlations between cognitive ability and bat and ball accuracy

	Bat and ball trial number							
	1	2	3	4	5	6	7	8
Initial response	−0.31	0.18	0.38	0.10	0.20	0.32	0.14	0.31
Final response	0.24	0.32	0.39	0.30	0.36	0.21	0.36	0.36

Following Raoelison et al. (2020), we excluded participants if they reported familiarity with the bat and ball problem, missed the response deadline, or failed to recall the mnemonic load. After those exclusions, the sample sizes used to compute these correlations were about 55 for each trial.

We doubt that selecting the correct answer from a small set of provided options after repeated exposure to variants of the same problem represents anything resembling an intuitive solution to the bat & ball problem. Given the instructions which emphasize two kinds of answers and repeated exposure to isomorphs of the standard problem, we suspect, instead, that respondents either eventually recognize that their intuition may be incorrect (and thereby start choosing a counter-intuitive answer from the provided list) or learn to apply a problem-specific shortcut they eventually discover, such as dividing the difference between the two numbers by two.[37]

Though we reject Bago and De Neys (2019) claim that an appreciable fraction of respondents can solve *this* problem without engaging in substantial deliberation, we take no issue with Bago and De Neys (2017) broader claim that *many* reasoning problems contain multiple competing principles, that more than one of them can *sometimes* be quickly apprehended, and that this can create conflict which reduces confidence in the more dominant intuition (if the problem states almost everyone in the sample *is* a lawyer, Bill likely is too, even though he sounds more like an engineer). We also support (and, indeed, provide further evidence for) their proposal of rapid, nearly unconscious monitoring of the quality of quickly generated candidate responses.

As a final note regarding intuitive and deliberative responding, it seems important to distinguish the claim that 5 may be an intuitive response from the distinct, but related (?) claim, that some who ultimately solve the bat and ball problem never entertained 10 cents as a potential response. For instance, Szaszi and co-authors (2017) asked 219 respondents to solve the bat & ball problem out loud. Of the 38 who solved it only 14 *explicitly* mentioned the 10-cent intuition. While we assume this substantially underestimates the fraction who *computed* or *considered* that value, we agree that some who possess the ability to solve the problem may not seriously entertain 10 cents as a potential response, perhaps because they (a) immediately encode it as an algebra problem and start doing the math, (b) disbelieve they'd be asked to merely subtract 100 from 110, (c) notice that the second statement does *not* simply say that the bat costs $1, which it would if subtraction were the only required operation, or (d) somehow intuitively appreciate the principle that one can create a difference of n units between two things by subtracting n/2 units from one thing and adding it to the other (if Andrew gives Shane $5, the *difference* in their wealth has increased by $10).

[33] Bago and De Neys (2019) note that time may elapse between a thought and an overt response, and speculate that if time limits are not imposed, response latencies may include additional cogitation ("double-checking") of accurate intuitions that were produced more rapidly. However, this could not explain why respondents who answer 5 cents take much longer to respond (57 s) than those who give incorrect answers (28 s). Why would respondents choose to cogitate longer upon an *accurate* intuition than an *inaccurate* one? – especially considering that accurate responses are held with even greater confidence than inaccurate ones (as discussed in Appendix G). We favor a more traditional account: people take longer to respond when they *need* more time to generate that response.

[34] Further, if respondents given more time to reflect typically repeat correct responses but only sometimes revise incorrect ones, a claim that cognitive ability correlates *more* strongly with the percentage of trials initially answered correctly than with the percentage of trials for which respondents initially failed but later succeeded seems akin to a claim that making both free throws (vs. fewer) will correlate more strongly with basketball ability than missing the first and making the second (vs. missing both or making both). For *any* test that is monotonically related to some criterion variable, achieving a perfect score (vs. scoring lower) will always correlate more highly with the criterion than achieving any intermediate score (vs. scoring lower or higher).

[35] Item order was not recorded in their first (and only other) experiment.

[36] On the first trial, the initial response to the other two tasks that Raoelison et al. (2020) investigated (a base rate problem and a belief-bias syllogism) also had no correlation with their measure of cognitive ability (rs = −0.06 and 0.07).

[37] In all of these multi-trial studies, the standard problems (which involve the key words "more than") were intermixed with "lite" versions (which lack these words, and for which subtracting the smaller number from the larger one yields the correct result). During this long set of intermixed trials, a subset of respondents eventually acquired the ability to solve the standard problems. However, performance on the lite items remained near ceiling throughout, which means that these participants learned to distinguish these two types of problems, rather than just dividing the difference by 2 for all problems encountered (which would yield correct answers for the standard problems but *incorrect* answers for the "lite" problems).

Appendix M. The careless and the hopeless

Some of our within-subject *Hint* experiments included questions besides the Bat and Ball problem: six items from Raven's Advanced Progressive Matrices, which is widely upheld as a measure of general intelligence (Jensen, 1998) and Tversky and Kahneman's (1983) "Linda question" (which also plausibly measures the ability to resist an intuition).

Some of these studies also included the other two items from Frederick's (2005) Cognitive Reflection Test and employed the same within-subject *Hint* procedure. For "Widgets",[38] we told subjects that the answer was not 100. For "Lilypads",[39] we told them it was not 24. As shown in the tables below, the result for the bat & ball item reported in the main text replicates for these two items as well: in both cases, the largest gap in Raven's scores is between the *careless* and the *hopeless*, whereas the largest gap in the "Linda" problem is between the *reflective* and the *careless*.[40]

Table M1
Raven's and Linda performance by Widgets response $_{\text{Number of observations}}$

	Raven's score (out of 6)	Avoiding Conjunction Fallacy
Reflective (5 on first try)	3.4 $_{668}$	39% $_{301}$
Careless (100, but revised to 5)	3.3 $_{227}$	22% $_{117}$
Hopeless (100, and never got 5)	2.5 $_{741}$	20% $_{341}$

Table M2
Raven's and Linda performance by Lilypads response $_{\text{Number of observations}}$

	Raven's score (out of 6)	Avoiding Conjunction Fallacy
Reflective (47 on first try)	3.5 $_{850}$	34% $_{418}$
Careless (24, but revised to 47)	3.1 $_{114}$	20% $_{56}$
Hopeless (24, and never got 47)	2.3 $_{686}$	18% $_{311}$

The Hint manipulation enables us to distinguish the ability to catch the intuitive error on one's own from the ability to perform the required math once the error has been pointed out. Table M3 reports how solving these items (either without or with hints) predicts performance on a second reasoning task (the Raven's items or the "Linda" problem). For all three CRT items, the hint *strengthened* the relation with Raven's scores, but *weakened* the relation with solving the Linda problem (i.e., avoiding the conjunction fallacy). If the items are aggregated, both of these "opposing" effects are significant.[41]

Table M3
Relation between performance on CRT items (before & after hint), and performance on two other reasoning tasks (six Raven's matrices & the Linda problem).

Dependent Variable: Raven's Score (0 to 6)			
Independent Variable:	Bat and Ball $_{N = 1947}$	Widgets $_{N = 1948}$	Lilypads $_{N = 1947}$
Constant	2.35 $_{0.05}$	2.41 $_{0.05}$	2.27 $_{0.05}$
Solved item before hint	0.28 $_{0.10}$	0.11 $_{0.12}$	0.44 $_{0.14}$
Solved item after hint	0.85 $_{0.09}$	0.88 $_{0.11}$	0.79 $_{0.14}$
Wald-test of difference between before and after	z = 3.3	z = 3.6	z = 1.3

Dependent variable: committing Conjunction Fallacy (0) or avoiding it (1)			
Independent Variable:	Bat and Ball $_{N = 1386}$	Widgets $_{N = 929}$	Lilypads $_{N = 929}$
Constant	0.19 $_{0.02}$	0.21 $_{0.02}$	0.20 $_{0.02}$
Solved item before hint	0.12 $_{0.03}$	0.15 $_{0.05}$	0.15 $_{0.06}$
Solved item after hint	0.02 $_{0.03}$	0.02 $_{0.04}$	-0.00 $_{0.06}$
Wald-test of difference between before and after	z = 1.9	z = 1.6	z = 1.4

Table reports OLS regression coefficients with standard errors as subscripts. Note that "Solved item after hint" equals 1 if the participant eventually solved the item, regardless of whether the solution was actually entered before or after the hint.

Correspondingly, Fig. M shows that those with higher cognitive abilities (operationalized by their Raven's scores) were not only more likely to solve the bat and ball problem initially (*reflective* responses) but also more likely to use the hint to correct their initial error (*careless* responses).

[38] If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets?

[39] In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

[40] Atypical answers are more common for the Widgets and Lilypads problem (17% and 15%, respectively) than for the bat and ball problem (5%).

[41] Predicting Linda: $B_{initial} = 0.27$, vs. $B_{final} = -0.01$, $z = 2.4$. Predicting Ravens: $B_{initial} = 0.42$ vs. $B_{final} = 1.22$,; $z = 2.8$.

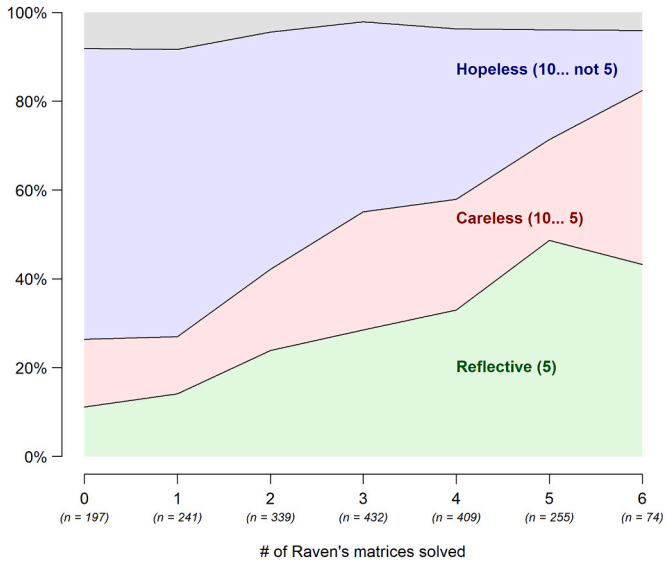


Fig. M. Distribution of bat and ball responses by performance on Raven's APM
The Reflective area indicates the percent of respondents initially answering correctly (5). The Careless area indicates the percent of respondents initially answering with the intuitive error (10), but later revising to 5 when told that 10 is wrong. The Hopeless area indicates the percent of respondents initially answering 10, but failing to revise to 5 when told that 10 is wrong. The unlabeled grey area indicates the percent of respondents who initially answered something other than 5 or 10.

References

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of experimental psychology: General, 136*(4), 569.

Attali, Y., & Bar-Hillel, M. (2020). The false allure of fast lures. *Judgment and Decision making, 15*(1), 93.

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition, 158*, 90–109.

Bago, B., & De Neys, W. (2019). The smart system 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning, 25*(3), 257–299.

Bago, B., Raoelison, M., & De Neys, W. (2019). Second-guess: Testing the specificity of error detection in the bat-and-ball problem. *Acta Psychologica, 193*, 214–228.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265–284.

Beattie, J., & Baron, J. (1988). Confirmation and matching biases in hypothesis testing. *The Quarterly Journal of Experimental Psychology, 40*(2), 269–297.

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods, 50*(5), 1953–1959.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. E. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about Knowing*. The MIT Press.

Boissin, E., Caparos, S., Raoelison, M., & De Neys, W. (2021). From bias to sound intuiting: Boosting correct intuitive reasoning. *Cognition, 211*, Article 104645.

Borghans, L., Meijers, H., & Ter Weel, B. (2008). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry, 46*(1), 2–12.

Bourgeois-Gironde, S., & Van Der Henst, J. B. (2009). How to open the door to system 2: Debiasing the bat-and-ball problem. In S. Watanabe, et al. (Eds.), *Rational animals, irrational humans* (pp. 235–252). Keio University.

Bruner, J. S., & Potter, M. C. (1964). Interference in visual recognition. *Science, 144* (3617), 424–425.

Burič, R., & Konrádová, Ľ. (2021). Mindware instantiation as a predictor of logical intuitions in the cognitive reflection test. *Studia Psychologica, 63*(2), 114–128.

Claidière, N., Trouche, E., & Mercier, H. (2017). Argumentation and the diffusion of counter-intuitive beliefs. *Journal of Experimental Psychology: General, 146*(7), 1052.

De Neys, W., & Bonnefon, J. F. (2013). The 'whys' and 'whens' of individual differences in thinking biases. *Trends in Cognitive Sciences, 17*(4), 172–178.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition, 106*(3), 1248–1299.

De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 208–216.

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.

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science, 19*(5), 483–489.

DeStefano, D., & LeFevre, J. A. (2004). The role of working memory in mental arithmetic. *European Journal of Cognitive Psychology, 16*(3), 353–386.

Enke, B., Gneezy, U., Hall, B., Martin, D., Nelidov, V., Offerman, T., & van de Ven, J. (2021). Cognitive biases: Mistakes or missing stakes? *The Review of Economics and Statistics*, 1–45.

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

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives, 19*(4), 25–42.

Hoover, J. D., & Healy, A. F. (2017). Algebraic reasoning and bat-and-ball problem variants: Solving isomorphic algebra first facilitates problem solving later. *Psychonomic Bulletin & Review, 24*(6), 1922–1928.

Hoover, J. D., & Healy, A. F. (2019). The bat-and-ball problem: Stronger evidence in support of a conscious error process. *Decision, 6*(4), 369.

Hoover, J. D., & Healy, A. F. (2021). The bat-and-ball problem: A word-problem debiasing approach. *Thinking & Reasoning*, 1–32.

Janssen, E. M., Raoelison, M., & de Neys, W. (2020). "You're wrong!": The impact of accuracy feedback on the bat-and-ball problem. *Acta Psychologica, 206*, Article 103042.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The doubting system 1: Evidence for automatic substitution sensitivity. *Acta Psychologica, 164*, 56–64.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81).

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In *The Cambridge handbook of thinking and reasoning* (pp. 267–293).

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237.

Keren, G. (2013). A tale of two systems: A scientific advance or a theoretical stone soup? Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science, 8* (3), 257–262.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science, 4*(6), 533–550.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97.

Lawson, M. A., Larrick, R. P., & Soll, J. B. (2020). Comparing fast thinking and slow thinking: The relative benefits of interventions, individual differences, and inferential rules. *Judgment and Decision Making, 15*(5).

Lawson, M. A., Larrick, R. P., & Soll, J. B. (2022). When and why people perform mindless math. *Judgment and Decision Making, 17*(6), 1208–1228.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs, 54*(6), i.

Margolis, H. (1987). *Patterns, thinking, and cognition: A theory of judgment*. University of Chicago Press.

Mastrogiorgio, A., & Petracca, E. (2014). Numerals as triggers of system 1 and system 2 in the 'bat and ball' problem. *Mind & Society, 13*(1), 135–148.

Mata, A. (2020). An easy fix for reasoning errors: Attention capturers improve reasoning performance. *Quarterly Journal of Experimental Psychology, 73*(10), 1695–1702.

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology, 105*(3), 353–373.

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: An attentional account of reasoning errors. *Psychonomic Bulletin & Review, 24*(6), 1980–1986.
Mata, A., Schubert, A. L., & Ferreira, M. B. (2014). The role of language comprehension in reasoning: How "good-enough" representations induce biases. *Cognition, 133*(2), 457–463.
Mayer, R. E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories, and templates. *Instructional Science, 10*(2), 135–175.
Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences, 22*(4), 280–293.
Meyer, A., Attali, Y., Bar-Hillel, M., Frederick, S., & Kahneman, D. (2023). *The cognitive reflection test is not "just" math: An adversarial collaboration* (Working Paper).
Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., … Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General, 144*(2), Article e16.
Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the cognitive reflection test. *Judgment and Decision making, 13*(3), 246.
Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215*(5109), 1519–1520.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175.
Otero, I., Salgado, S. F., & Moscoso, S. (2022). Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. *Intelligence., 90*, 1–13.
Patel, N., Baker, S. G., & Scherer, L. D. (2019). Evaluating the cognitive reflection test as a measure of intuition/reflection, numeracy, and insight problem solving, and the implications for understanding real-world judgments and beliefs. *Journal of Experimental Psychology: General, 148*(12), 2129.
Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods, 48*(1), 341–348.
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.
Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making, 29*(5), 453–469.
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.
Raoelison, M., Keime, M., & De Neys, W. (2021). Think slow, then fast: Does repeated deliberation boost correct intuitive responding? *Memory & Cognition*, 1–11.
Raoelison, M. T., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition, 204*, Article 104381.
Risen, J. L. (2016). Believing what we do not believe: Acquiescence to superstitious beliefs and other powerful intuitions. *Psychological Review, 123*(2), 182.
Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in god. *Journal of Experimental Psychology: General, 141*(3), 423.
Shweder, R. A. (1977). Likeness and likelihood in everyday thought: Magical thinking in judgments about personality. *Current Anthropology, 18*(4), 637–658.
Silva, J. D. S. (2005). *Se um taco e uma bola custam R $110, e o taco custa R $100 a mais do que a bola, quanto custa a bola?* (Doctoral dissertation).
Simon, G., Lubin, A., Houdé, O., & De Neys, W. (2015). Anterior cingulate cortex and intuitive bias detection during number conservation. *Cognitive Neuroscience, 6*(4), 158–168.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3.
Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the cognitive reflection test is stable across time. *Judgment and Decision making, 13*, 260–267.
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–665.
Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: Exploring the ways individuals solve the test. *Thinking & Reasoning*, 1–28.
Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107–140.
Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*(7), 1275.
Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the cognitive reflection test. *Cognition, 150*, 109–118.
Trémolière, B., & De Neys, W. (2014). When intuitions are helpful: Prior beliefs can support reasoning in the bat-and-ball problem. *Journal of Cognitive Psychology, 26*(4), 486–490.
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.
Tulving, E. (1985). How many memory systems are there? *American Psychologist, 40*(4), 385.
Tulving, E., Schacter, D. L., & Stark, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*(4), 336.
Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293.
Wason, P. C., & Evans, J. S. B. (1974). Dual processes in reasoning? *Cognition, 3*(2), 141–154.
Weiskrantz, L. (2009). Is blindsight just degraded normal vision? *Experimental Brain Research, 192*(3), 413–416.
Weiskrantz, L., Warrington, E. K., Sanders, M. D., & Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain, 97*, 709–728.