The Bat and Ball Problem

Andrew Meyer, Bob Spunt, & Shane Frederick

EXTREMELY ROUGH DRAFT

Abstract

The Bat and Ball problem has been upheld as a thin slice measure of an individual's disposition or ability to engage in reflective thought, and is now included as a covariate in many studies. Performance on it has been shown to correlate with intertemporal choice, risky choice, moral reasoning, strategic behavior, and belief in god. However, there is no account of why people miss the problem at such high rates -- why people conclude that the titular objects cost 10 cents and $1.00, despite the specification that their prices differ by $1.00.

In this paper, we propose a modified version of Kahneman & Frederick's (2002) attribute substitution hypothesis to explain the high error rates. In our view, respondents misread the problem and solve a simpler variant of it (to which their answers are correct). But even when they notice the discrepancy between their construal of the problem and the wording of the problem, they fail to realize that their answers do not also satisfy the stated constraints. They are able to maintain the beliefs that (a) their responses are $1.00 and $0.10 and (b) that their responses differ by a $1.00. We find evidence that exposure to vignettes simultaneously referencing the sum and difference of two objects momentarily reduces participants' ability to execute subtraction, which, although itself completely unexplained, largely explains participants' failure to recognize the contradiction between the answers they give and the difference they believe exists between them. We find that this illusion largely persists in the face of manipulations meant to decrease intuitive confidence, though it is markedly reduced when the value yielded by the attribute substitution no longer represents a plausible response to the question.

The “bat and ball” problem

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

Most people with a college degree – and nearly everyone without one – answers 10 cents (Frederick, 2005). Though one can readily devise other mathematical story problems that most people miss, the high error rates here are surprising, because this answer so plainly contradicts the second constraint: If the ball cost 10 cents and the bat costs $1.00 more, the bat, *itself*, would cost $1.10, and the two would sum to $1.20.

The simplicity and decisiveness of the disproof suggests that those who say 10 cents may not even have bothered to check whether that response satisfies the two stated conditions. Indeed Kahneman and Frederick (2002, 2005) cited this problem to show how readily intuitions are endorsed. This problem is now widely held up to illustrate the distinction between two types of reasoning processes: the reflexive, intuitive “System 1” processes that blurt out 10 cents, and the more cautious, reflective “System 2” processes, which intervene to reject this response.

To anyone who thinks 10 cents initially, then reflects a moment longer to see that 5 cents is the correct response, this story has the ring of truth. However, it is not an especially detailed portrayal of the relevant cognitive processes, and may be inaccurate in some important respects. For instance, though the problem has been cited as showing that respondents don’t bother to check their answers (Kahneman and Frederick 2005, p. 273), perhaps the 10 cent response is produced *despite* performing many of the appropriate checks.

Our goal in this paper is to provide an in-depth analysis of the “bat and ball problem,” so as to help clarify what it does and does not reveal. We intend this research to contribute to the larger discussion of intuitive errors and intuitive confidence (Simmons & Nelson, 2005; xxx; yyy). Our focus on this specific problem reflects its prominent position in discussions regarding dual process views of cognition, and the large number of studies that have used this problem as a putative measure of reflective tendencies (see Aaron’s table)..

The Substitution Hypothesis

Consider the problem below:

1. Bat and Ball lite

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

10 cents is the overwhelming response to this problem as well;[1] but in this case, it is correct. If you find yourself stopping here to revisit how the *original* problem was worded, you can appreciate how easy they are to confuse. Indeed, since $1.10 and $1.00 are explicitly mentioned in the standard problem, and the words "more than" suggest the computation of a difference, it may actually be difficult to *avoid* solving the simplified problem when reading the "standard" problem (which explains the high error rate, and the nearly universal tendency to *consider* 10 cents as a possible response, even among those who ultimately reject it).

That substitution predicts a specific belief about the cost of the bat. In fact, if participants are asked to specify the cost of *both* objects, nearly all who make the error conclude that the bat and ball cost $1.00 and $0.10, respectively (satisfying the first constraint, but violating the second), rather than $1.10 and $0.10 (which violates the first constraint, but satisfies the second). This remains true regardless of the order in which the constraints are presented (see Appendix A).

Subtracting $1.00 from $1.10 (bat and ball lite) is much easier than determining two values which both differ by $1.00 and sum to $1.10 (bat and ball). Interestingly, Frederick (2005) reported that participants who respond $0.10 actually think the bat and ball problem was easier than participants who respond $0.05.[2]

Our conjecture that respondents unwittingly substitute the easier problem for the harder one is, essentially, the *attribute substitution* hypothesis proposed by Kahneman and Frederick (2002, 2004, 2005). In the context of this problem, this hypothesis receives further support from work by Mayer (1981), who found that algebraic word problems assign values to *variables* (as in the lite version above) much more commonly than they assign values to *relations* (as in the standard problem). Moreover, Mayer (1982) showed that relational propositions are often misremembered as assignment propositions – precisely the substitution we posit. We test this directly next.

[1] We used Google surveys to administer the Bat & Ball problem and its simplified variant to representative samples of American web-browsers. Those who received the Bat & Ball problem (n = 1,096) solved it 8% of the time and said 10 cents 76% of the time whereas those who received the simplified variant (n = 1,055), where 10 cents was the solution, said 10 cents 87% of the time. Effectively, including the words "more than the ball" reduced 10 cent response from 87 to 76% while increasing 5 cent response from 0 to 8%.

[2] Specifically, Frederick (2005) asked participant to estimate the percentage of other participants who got the problem right. He found that 5 cent respondents estimated lower percentages than 10 cent respondents. We replicated that relation (). We also asked a group to directly estimate the problem's difficulty. The relation between their bat and ball responses and their easiness ratings supported our interpretation of the other participant performance judgment ().

Study 1 (Recall of Problem Wording)

We presented the Bat & Ball problem to 971 participants from mTurk and eLab.[3,4] After participants entered their response, the text disappeared, and they were asked to reproduce the problem from memory. A coder (blind to the participants' answer to the bat & ball problem and to our hypothesis) classified the recreated problem as "correct" if $1.00 referenced the *difference* in cost between the bat and ball, as "lite" if $1.00 referenced the price of the bat, and as "idiosyncratic" if the problem recreated from memory contained other error or errors (For example: **"**there is a difference of 10 cents between price of bat and ball bat is $ 1 more than ball , how much is the price of ball?" or "No").

Results & Discussion:

The problem respondents remember answering

Response:	**standard**	**lite**
5 cents n = 158	94	0
10 cents n = 397	61	23
other n = 60	43	7

As predicted, those who said 10 cents committed the hypothesized mnemonic error more frequently than those who said 5 cents (23% vs. 0%; $z(553) = 6.41$; $p < .001$) or those who gave some atypical response (23% vs. 7%; $z(455) = 2.66$; $p = .008$). Those who recalled it as the "lite" version almost always said 10 cents (and those who solved the problem could almost always correctly recall its structure).[5] That said, 61% of those who said 10 cents were able to reproduce the problem correctly. Thus, awareness (or at least memory) of the problem's actual wording does not guarantee that respondents attend to the four critical words that differentiate the two problems (i.e., "more than the ball"). We suspect that many respondents perform the posited substitution, solve that version problem, and notice some extra words, which they are able to recall, but which are not a sufficient cue to have them re-examine their interpretation of the problem.

We interpret these results as providing *some* support for the substitution hypothesis, though it is possible that some other factor, besides the substitution, causes both the 10 cent response and mnemonic error. It is also possible that the 10 cent response itself creates the faulty memory -- a form of confirmation bias (Nickerson, 1998). In the following experiments, we examined the strength of cues required to inhibit the posited substitution.

[3] We excluded 22 participants who quit the survey before arriving at the mnemonic task. And we excluded an additional 334 people who said that they had seen the problem before.

[4] This analysis includes data from control conditions and ineffective manipulations (including the font of the problem's text and the names of the objects) from a number of studies reported later in this paper.

[5] We obtained similar results in a separate study that replaced the free recall task with a recognition task. Participants answered the bat and ball problem, moved on to the next page where they were presented with both the regular bat and ball problem and bat and ball lite, and were asked which of the two problems they had previously answered. Among 10 cent respondents, 24% chose the lite problem (n = 225). Among 5 cent respondents, 0% chose the lite problem (n = 164). And among respondents making other errors, 19% chose the lite version (n = 27).

2. Attempts to inhibit the Substitution

Study 2 (Emphasizing that $1.00 describes a *relation*, not a *price*)

Following our conjecture that many intuitive errors arise from participants' failure to appreciate that $1.00 refers to a *difference* in price, we emphasized the words "more than the ball." In the first, 330 respondents on eLab were randomly assigned to one of three conditions: a standard problem control, a condition in which the words "more than the ball" were bolded, and a *Contrast* condition in which respondents encountered both the lite and regular problem, in that order.[6] We reasoned that conversational norms against redundant questions and the visually apparent difference in the number of words required to express the problem might emphasize the critical phrase more effectively than bolding it.

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

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

***CONTRAST* CONDITION**
A bat and a ball cost $1.10 in total. The bat costs a dollar.
How much does the ball cost? ______

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

Results & Discussion

To our surprise, neither manipulation markedly affected performance. In the control condition, 29% of respondents solved it, compared with 24% in the bold condition, and 35% in the Contrast condition.[7] We followed this up with a large sample test of the bolding manipulation. It replicated these null results (See Appendix B) In two other studies, we manipulated the problem's wording in other ways in an attempt to inhibit the substitution we expected. These manipulations were also unsuccessful (see Appendix C).

Discussion

These data appear to weigh against the substitution hypothesis. But we can think of no better single explanation for why the majority of people say that the ball costs $0.10; for why nearly all of those $0.10 respondents say that the bat costs $1.00; for why those $0.10 respondents seem to think that they answered an easier question than did other respondents; and for why the $0.10

[6] We omit 90 of those participants from our analysis because they reported having seen the question before.

[7] FOOTNOTE discussing the Google binary replication, with reference to an appendix in which it is spelled out in more detail.

respondents are particularly likely to forget the words that differentiate this problem from its hypothesized substitute.

These data, in conjunction with the only partial success of the mnemonic tests from study 1, push us toward a weaker version of the substitution hypothesis. Even people who succumb to the substitution are probably aware of the words, “more than the ball.” The majority recall them. And at least some probably recognize that they are important. Although quite confident in the accuracy of their response, 10 cent respondents are less confident than 5 cent respondents (respectively, 73% and 83% estimated a 100% chance of being correct[8]).

Perhaps respondents start out by mistaking this problem for the simplified variant, notice some discrepancy, but already possessing a strong candidate, simply accept it. In essence, the substitution is a largely ballistic process. Once people make the mistake, it gains sufficient momentum to overcome many doubts, including those raised by text emphasizing that $1.00 refers to a difference, rather than the price of the bat. However, we assume it will not be able to overwhelm all possible doubts.

In the next series of studies we attempted to make more salient the actual difference between the bat and ball’s prices that was implied by the participant’s response.

[8] After answering the question and while their response and the question text were still on the screen, 433 5 cent respondents and 774 10 cent respondents specified the probability that their response was correct on an 11 point scale from 0 to 100%. 104 respondents who gave answers other than 5 or 10 were also asked. Only 46% of them were 100% confident in their response. Similarly, Wim De Neys (2013) reported that erroneous 10 cent respondents to this problem are less confident than accurate 10 cent respondents to its simplified variant.

3. Salience of the Difference between the Bat's and Ball's Prices

Study 3a (Specifying the Bat's price too)

We randomly assigned XXX[9] mTurk participants to answer either the standard bat and ball question or a version of it that asked about both the ball's price and the bat's.

CONTROL
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? $______

***BAT TOO* CONDITION**
A bat and a ball cost $1.10 in total. The ball costs $1.00 less than the bat.
How much does the ball cost? $______
How much does the bat cost? $______

Results:
To our surprise asking participants to specify the bat's price along with the ball's neither increased performance (34% vs. 27% correct) nor decreased the 10 cent error (54 vs. 51%).

In the next experiment, we tried to simultaneously make more obvious both the 90 cent difference between the 10 cent and $1.00 responses and the $1.00 difference between the 5 cent and $1.05 responses.

Study 3b (Both Items Together)

In this experiment, response was binary, rather than open-ended. We used Google Surveys to randomly assign 2,006 web-browsers to one of two conditions. In the control, participants were simply asked to choose between a $5 and $10 price for the ball. In the *Both Prices Together* condition, participants were asked for the prices of both the bat and the ball individually. They responded by choosing between "$105 and $5" and "$100 and $10."

CONTROL
A bat and a ball cost $110 in total. The bat costs $100 more than the ball.
How much does the ball cost? $5 OR $10

***BOTH PRICES TOGETHER* CONDITION**
A bat and a ball cost $1.10 in total. The ball costs $1.00 less than the bat.
How much do the bat and ball cost individually? $105 and $5 OR $100 and $10

Results:

Performance did improve when both prices were placed together on a response option (20% vs. 33% correct, z(2,004) = X.XX, p < .001). However, the majority still said $100 and $10, either failing to notice that those two prices ought to differ by $100, or failing to notice that those two prices do not differ by $100.

[9] We excluded XXX participants who said that they had seen the problem before.

In the next experiments, we had participants solve the bat and ball problem before directly asking them about the price difference between their bat and ball responses.

Studies 3c & d (Do your answers differ by $1.00?)

In two experiments, we randomly assigned 354[10]participants to one of two conditions. We ran the first experiment on mTurk and the second on a commuter ferry between Long Island and Connecticut. In both experiments, the control condition described a bat and a ball that cost $1.00 and $0.10 respectively and a *confirmation* condition that presented participants with the bat and ball problem and asked them to solve for the prices of both the bat and the ball. In the control, participants were asked "with those prices, does the bat cost $1.00 more than the ball?" In the *confirmation* condition participants were asked "is your bat answer $1.00 more than your ball answer?" In the experiment that we ran on mTurk, we included an additional question about the sum of the two prices before the focal question about the difference between them.

CONTROL
A bat costs $1.00 and a ball costs $0.10.

With those prices, do the bat and ball cost $1.10 in total? Yes OR No[11]
With those prices, does the bat cost $1.00 more than the ball? Yes OR No

***CONFIRMATION* CONDITION**
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? $____

Do your two answers sum to $1.10? Yes OR No[12]
Is your "bat" answer $1.00 more than your "ball" answer? Yes OR No

Results
For each condition and experiment, the table below presents the overall percentages erroneously concluding that two values differed by $1.00 followed by that percentage of only those who considered a $1.00 bat and a $0.10 ball, either because those were the numbers presented to them (in the control) or because those were the answers they generated (in the *confirmation* condition).

% erroneously concluding that a pair of prices differed by $1.00…
Overall [Of those who considered $1.00 and $0.10]

Condition:	**C. mTurk** *N* = 90	**D. Ferry** *N* = 133
Control	37 [37]	6 [6]
Confirmation	54 [86]	52 [74]

In both experiments, the overwhelming majority of participants (86 and 74% respectively) who said that that the ball cost $0.10 and the bat cost $1.00 maintained that those answers differed by $1.00. Overall, presenting the bat and ball problem for consideration dramatically increased the

[10] 131 participants are omitted from the analysis because they reported having seen the bat and ball problem before.
[11] This question was only present in the mTurk experiment.
[12] This question was only present in the mTurk experiment.

rate at which participants erroneously implied that $0.10 and $1.00 differed by $1.00 (overall 19 vs. 53%; *z*(221) = 5.14; *p* < .001).

Discussion:

The most obvious explanation is presentational: participants were unwilling to admit to the experimenter that their answers were wrong. Although we assume that that is a factor, we don’t think it is the only factor. We wondered if intuiting that the prices were $1.00 and $0.10 inhibited the ability to perform operations that would contradict that intuition, an intuitional shielding of sorts.

4. Inhibiting Contradictory Operations

To determine whether considering the bat and ball problem inhibited the operations required to check the 10 cent response to it, we measured the effect of reading about two objects that cost $110 in total and differ in cost by $100 on participants' ability to determine the difference between 100 and 10.

Study 4a-c (Al & Bob)

We used Google surveys to randomly assign 14,189 web-browsers to read one of three vignettes about a man named Al who bought a phone and a pen. They were:

$240 PHONE & PEN **CONTROL**
Al paid $240 for a phone & pen. He paid $100 more for his phone than his pen.

$110 PHONE **CONTROL**
Al paid $110 for a phone. He paid $100 more for his phone than his pen.

$110 PHONE & PEN **CONDITION**
Al paid $110 for a phone & pen. He paid $100 more for his phone than his pen.

In experiment A participants were then asked "Bob paid $100 for a phone & $10 for a pen. How much more did Bob pay for his phone than his pen?" and responded by choosing between $90 and $100.

Experiment B was identical, except rather than offering participants a choice between $90 and $100, it asked them to type a number into an open-ended response blank.

In experiment C the question about the difference in price between Bob's items was replaced with the simpler one: "What is 100 minus 10?, " to which participants responded by typing a number into an open-ended response blank.

In the *$110 for a phone & pen* condition we hypothesized that participants who read the Al vignette would automatically think that Al's phone and pen cost $100 and $10, despite being aware that those two prices should differ by $100. We predicted that that conflict would reduce participants' ability to produce the difference between 100 and 10 when subsequently asked about it. Both controls were meant to remove that specific conflict while matching the *$110 for a phone & pen* condition as closely as possible in all other ways.

Results:

For each experiment and condition, the table below presents the percentages responding that Bob's $100 phone cost $90 more than his $10 pen or that 100 was 90 more than 10. The subscripts are the percentages concluding that the two numbers instead differed by 100.

% responding 90 $_{\text{\% responding 100}}$

Condition:	**A. Binary Bob** $_{N = 5{,}028}$	**B. Open-ended Bob** $_{N = 3{,}110}$	**C. Open-ended** $_{N = 6{,}051}$
$240 phone & pen	78	59 $_{5}$	90 $_{0.4}$
$110 phone	74	55 $_{10}$	89 $_{0.7}$
$110 phone & pen	71	48 $_{15}$	87 $_{1.0}$

In all experiments, participants were worst at determining the difference between 100 and 10 in the *$110 for a phone & pen* condition (compared to the *$110 phone* condition: experiment a: $z(4{,}023) = 2.16$, $p = .031$; experiment b: $z(2{,}070) = 2.94$, $p = .003$; experiment c: $z(4{,}037) = 2.28$, $p = .023$; compared to the *$240 phone & pen* condition: experiment a: $z(3{,}008) = 3.80$, $p < .001$; experiment b: $z(2{,}077) = 5.29$, $p < .001$; experiment c: $z(4{,}027) = 3.29$, $p = .001$).

While we predicted that that specific subtraction would be inhibited as some kind of defense against simultaneous contradictory belief (Sloman, 1996), that prediction implies that people already know, at some level, that that specific operation will threaten their belief. We wind up with a theory that requires that people already know the result of the subtraction in order for there to be any reason that they not be able to complete that subtraction, a potentially serious conceptual flaw.

In experiment 4d, we varied the calculation following the Al & Bob manipulation to determine which calculations would be affected and which would not.

Study 4d (Inhibiting Subtraction)

We used Google surveys to randomly assign 10,246 participants to one of 10 conditions in a 2x5 design. Participants either were exposed to the *$110 phone* or the *$110 phone & pen* vignette from experiments 7a-c before answering one of five questions about Bob's phone & pen:

CRITICAL SUBTRACTION
Bob paid $100 for a phone & $10 for a pen. How much more did Bob pay for his phone than his pen?

SUBTRACTION WITH $90 ANSWER
Bob paid $140 for a phone & $50 for a pen. How much more did Bob pay for his phone than his pen?

UNRELATED SUBTRACTION
Bob paid $140 for a phone & $20 for a pen. How much more did Bob pay for his phone than his pen?

CRITICAL ADDITION
Bob paid $100 for a phone & $10 for a pen. How much did Bob pay in total?

UNRELATED ADDITION
Bob paid $140 for a phone & $20 for a pen. How much did Bob pay in total?

Results:

The table below presents the percentages responding correctly for each required operation and Al vignette condition.

% responding correctly

Condition:	**Critical Subtraction** *N* = 2,023	**Subtraction with $90 Answer** *N* = 2,052	**Unrelated Subtraction** *N* = 2,037	**Critical Addition** *N* = 2,111	**Unrelated Addition** *N* = 2,023
$110 phone	57	64	61	76	74
$110 phone & pen	48	59	55	79	75

The decrease in solution of the *Critical Subtraction* in the *$110 phone & pen* condition replicated experiment 7b ($z(2,021) = -3.94$, $p < .001$). Additionally, we found similar deleterious effects on solution of the other two subtraction problems ($z(2,050) = -2.03$, $p = .042$ and $z(2,035) = -2.79$, $p = .005$). But neither addition problem was affected.

Discussion:

Considering two objects that sum to $110 and differ by $100 appears to inhibit subtraction. We don't know why that operation would be inhibited. But its inhibition would explain why people fail to notice that their 10 cent and $1.00 answers to the bat and ball question only differ by 90 cents.

In the next section, we attempted to manipulate confidence in intuition in a number of ways. In keeping with the literature on the subject, we predicted that various warning signals in the environment would shift people into a more reflective mindset, in turn reducing the rate of intuitive error and improving performance.

5. Confidence in Intuition

Alter and colleagues (2007) reported that performance on Frederick's (2005) cognitive reflection test (which includes the bat and ball problem) improved when it was merely printed in dysfluent font. They reasoned that people misattributed the difficulty of reading the problem to the problem itself, causing them to doubt their initial intuitions and go on to reason more carefully.

Studies 5a-d (Dysfluent font)

We report data collected by three independent research groups who manipulated the fluency of the font in which the bat and ball question was printed or displayed on the computer screen. Alter and colleagues (2007) reported data from 40 Princeton University undergrads. Thompson and colleagues (2012) reported data from 368 Saskatchewan University undergrads. And we collected data from 403 mTurk participants[13] and from 2,006 randomly selected web-browsers. In each sample, about half of the participants received the problem in a normal font, and half received it in a font that was difficult to read.

Results:

The table below displays the solution rates and percentages making the intuitive error (as a subscript).

$\%\ 5\text{ cents}_{\%\ 10\text{ cents}}$

Condition:	**A. Princeton**[14] $_{N = 40}$	**B. U of S**[15] $_{N = 368}$	**C. mTurk** $_{N = 266}$	**D. Google**[16] $_{N = 2{,}006}$
Control	80_{20}	35	28_{59}	19
Dysfluent	75_{20}	21	21_{59}	22

Overall, there was no effect of font on solution rates. In both conditions 23% of participants got the problem right. Only the Google data show any positive effect of dysfluent font. But it is binary response. So we interpret that slight improvement as a slight shift toward random response when the text is harder to read.[17]

There are other subtle ways to make the problem *feel* harder. You could replace the bat and ball with nonsense words, for example, a "clabor" and "plonket." We included "clabor and plonket" conditions in both our Google and mTurk experiments. In the Google data, the solution rate increased from the control condition's 19% to 25% (n = 1,017). But in the mTurk data, it

[13] We exclude 137 of these participants from analysis because they reported having seen the problem before.

[14] These data come from experiment 1 of Alter et al (2007).

[15] These data come from experiments 1 and 3 of Thompson et al (2012). We omit percentages committing the 10 cent error because Thompson did not provide us with those data.

[16] We omit percentages committing the 10 cent error because these are binary responses; percentages committing the 10 cent errors are the complement of percentages responding 5 cents.

[17] In Alter et al's data, there was a huge improvement in performance on the widget problem (20% to 80%) and no effect on the lilypad problem or the bat and ball problem (90% vs. 90% and 80% vs. 75%). However, we have no reason to believe that there is anything special about the widgets problem. Thompson et al (2012) attempted to replicate Alter's results with all three items and found no improvement in performance with dysfluent font on any of them.

decreased from 28% to 19% (n = 129). The difference between the two is probably the response format.

In another experiment, we put each constraint on its own row and numbered it so that the problem would look more “mathy.”

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

How much does the ball cost? 5 cents OR 10 cents

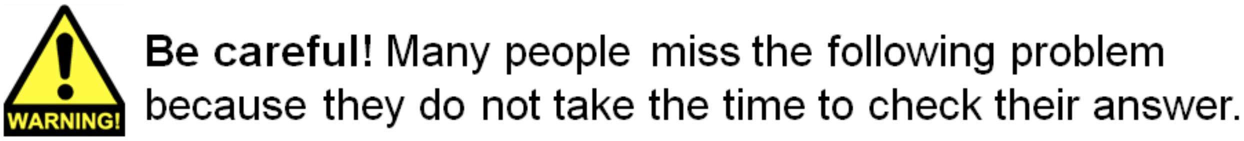
But that didn’t help performance at all either. If anything, it lowered it slightly, from 19% to 17% (z(2,0XX) = X.XX, p =).

Although discordant with previous results, the fact that these subtle manipulations of metacognitive difficulty had no effect on solution rates does not mean that confidence in the 10 cent intuition cannot be manipulated to some salutary effect. The simplest, most direct way we could think of to make people doubt their initial intuition was to tell them that the problem was tricky and warn them that they should check their answers.

Studies 5e-h (Explicit Warnings)

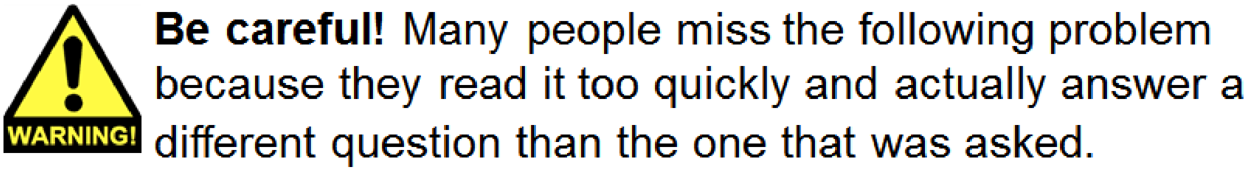
Across four similar studies in four populations, a total of 1,723 participants[18] received the Bat & Ball question, sometimes as part of the 3-item Cognitive Reflection Test (Frederick, 2005). Participants were randomly assigned to receive either just the bat and ball problem, or the bat & ball problem preceded by one of three warnings against relying on an initial intuition. As shown below, all three warnings urged caution, but differed in what was emphasized. The *computation* warning entreated participants to "check their answer." The *comprehension* warning cautioned against misreading the problem. And the *constraint* warning specifically asked participants to check that their answer satisfies both statements in the problem.

***COMPUTATION* WARNING**



Be careful! Many people miss the following problem because they do not take the time to check their answer.

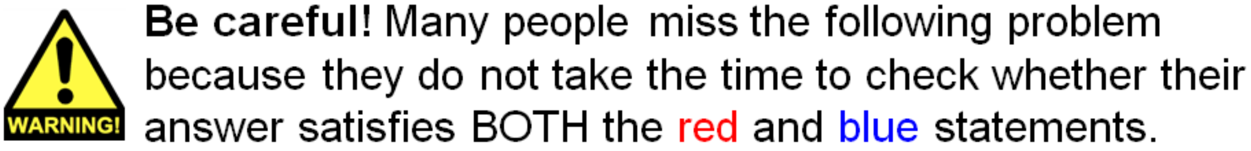
***COMPREHENSION* WARNING**



Be careful! Many people miss the following problem because they read it too quickly and actually answer a different question than the one that was asked.

[18] We exclude 361 of these participants from analysis because they reported having seen the problem before.

***CONSTRAINT* WARNING**



WARNING!

Be careful! Many people miss the following problem because they do not take the time to check whether their answer satisfies BOTH the red and blue statements.

[19]

We conducted two paper and pencil studies with 282 and 241 students and two online studies with 770 and 431 participants.

Results:

The table below displays the solution rates and percentages making the intuitive error (as a subscript).

% 5 cents $_{\%\ 10\ \text{cents}}$

Condition:	**E. UCLA** $_{N = 282}$	**F. Yale** $_{N = 241}$	**G. eLab**[20] $_{N = 607}$	**H. mTurk** $_{N = 238}$
Control	41 $_{58}$	79 $_{19}$	38 $_{51}$	34 $_{54}$
Computation	49 $_{48}$	84 $_{14}$	48 $_{46}$	25 $_{69}$
Comprehension	58 $_{41}$	91 $_{4}$	39 $_{52}$	44 $_{53}$
Constraint	--	--	34 $_{58}$	38 $_{52}$

Overall, performance was only slightly better with a warning than without one, (45% vs. 50% correct, $z(1{,}363) = 1.95$, $p = .051$),[21] a surprisingly small effect of a rather heavy-handed manipulation.

Finally, we report the results of a much stronger manipulation of confidence in the 10 cent intuition: simply telling participants that it is wrong.

Studies 5i - n (HINT: It's not 10 cents…)

Across six similar studies in four populations, a total of 2,619 participants[22] received the Bat & Ball question, sometimes as part of the 3-item Cognitive Reflection Test (Frederick, 2005). In four of those studies, participants were randomly assigned to either the *control* condition (just the bat & ball question) or a *hint* condition in which the words "HINT: 10 cents is <u>not</u> the answer." were printed to the right of the blank in which their response to the bat & ball question was entered. Those studies included 551 students at UCLA, 275 students at Yale, and two online studies with 766 and 533 participants.

[19] In this warning condition, the problem's first sentence "A bat and a ball cost $1.10 in total" was colored red, and the problem's second sentence, "the bat costs $1.00 more than the ball" was colored blue.

[20] In this study, half of the participants in each warning condition saw the normal second statement of the problem ("The bat costs $1.00 more than the ball.") and half saw this version: "The ball costs $1.00 less than the bat." That manipulation did not interact with the warning manipulation, and in the results above, we collapse across it. Note that these are the same data reported in experiment 2d. There, we collapsed across warnings to report the effect of the wording. Here, we collapse across wording to report the effect of the warning manipulation.

[21] All overall means and differences are estimated after controlling for study to study performance differences.

[22] We exclude 495 of these participants from analysis because they reported having seen the problem before.

In two additional online studies, 253 and 241 participants were placed in a within-subject design. They first entered their answer to the standard bat & ball question. After submitting that answer, they were given the Hint and an opportunity to revise their answer.

Results:

The table below displays the solution rates and percentages making the intuitive error (as a subscript).

% 5 cents $_{\%\ 10\ \text{cents}}$

	Between-Subject				**Within-Subject**	
Condition:	**I. UCLA**[23] $_{N = 551}$	**J. Yale** $_{N = 275}$	**K. eLab** $_{N = 387}$	**L. mTurk** $_{N = 535}$	**M. eLab** $_{N = 190}$	**N. mTurk** $_{N = 186}$
Control	42_{56}	65_{31}	33_{53}	38_{54}	27_{59}	35_{60}
With Hint	64_{6}	82_{3}	49_{13}	67_{20}	48_{17}	68_{16}

Since 10 cents is, by far, the most common error, we reckoned that invalidating that response would be an extremely effective "hint." It was fairly effective. The hint elevated performance in all studies (all *p*s < .005). Overall, solution rates went from 39% in the control to 62% with the Hint.

However, a significant minority of participants (13%) maintained 10 cents despite our hint. In the within-subject experiments, 61 participants maintained 10 cents as their answer. Of them, 46 simply did not change their response and 15 changed from one form of 10 cents to another – for example, from a decimal (.1) to a whole number (10). Although we meant to repudiate the content of their 10 cent response, some participants seem to have assumed that we were repudiating the form of their 10 cent response, suggesting that our Hint that the ball was not 10 cents was insufficient to overcome their belief that it was.[24]

Discussion

….

In the next section, we manipulated the salience of the correct answer. We predicted that encouraging participants to consider the correct response would cause them to reject the intuitive error in its favor.

[23] Not all hints are so helpful. In another condition within this population, we invalidated 20 cents. That hint did not aid solution, leaving its rate essentially unchanged at 43%.

[24] Rather than expect the hint to bring solution rates to 100%, a more lenient test of Frederick's conjecture might merely ask that all participants who change their answer from 10 cents, arrive at 5 cents. In fact, across all six experiments, the increase in 5 cent response is 57% of the decrease in 10 cent response, suggesting that about 57% of people who changed their answer as a result of the hint could in fact solve the problem.

6. Salience of the Correct Answer

We compared the standard open-ended response format to a choice between two options: 5 cents and 10 cents. We reasoned that presenting 5 cents to participants as one of two response options would cause many to consider it and realize that it was correct.

Study 6a (open-ended vs. binary response)

We collapse across four open-ended control conditions from experiments that we ran on Google surveys to estimate the open-ended solution rate, and across four binary response control conditions from experiments that we ran on Google surveys to estimate the binary response solution rate. That gives us 4,006 participants responding in the open-ended format and 4,522 in the binary response format.

OPEN-ENDED **FORMAT**
A bat and a ball cost $110 in total. The bat costs $100 more than the ball.
How much does the ball cost? $_____

BINARY-RESPONSE **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? 5 cents OR 10 cents

Results

With the open-ended format, 10% responded 5 cents and 77% responded 10 cents. With the binary choice format, 19% chose 5 and 81%, chose 10. It's probably misleading to think of the binary choice format as increasing the solution rate by 9%. You could think about the binary choice as redistributing the 13% (100-(10+77)) non-5 and non-10 responses from the open-ended format. If they were re-distributed randomly between the two options, you would expect about 16.5% (10+13/2) 5 cent responses, a bit less than the 19% that we observe (z(8,526)=2.6X, p =)[25], suggesting that the mere presence of the 5 cent response option caused a few people to solve the problem.[26]

In the next experiment, we again attempted to manipulate consideration of the 5 cent solution, but this time, less subtly.

[25] This is a two sample test, assuming sampling error in both the observed choice share and theoretical baseline.

[26] This comparison is confounded by a detail that is probably conservative to its conclusion. All of the binary responses were between a 5 and 10 cent ball. But three quarters of the open ended responses were to a version of the problem that we now prefer: one in which the $1.10 total and $1.00 difference are replaced by a $110 total and a $100 difference. In fact, that matters a little. In the open-ended format, the solution rate is 7% when there are decimals and 11% when there are not, p = .004. (Note that we code 5 AND .05 as solutions in both conditions, eliminating the most obvious effect of the decimal.) If we exclude the whole number bat & ball observations, we estimate the 5 cent choice share in the binary response format to be about 3 standard errors above the baseline that we would expect had non-5 / non-10 responses been randomly distributed between those two options.

Study 6b (Consider whether the answer could be $5.)

Using Google Surveys, we randomly assigned 2,003 web-browsers to one of two conditions. In the control condition, they simply answered the bat and ball problem. In the “consider 5 cents” condition, the words “before responding, consider whether 5 cents could be the answer,” appeared between the question and the answer blank.

CONTROL
A bat and a ball cost $110 in total. The bat costs $100 more than the ball. How much does the ball cost?

$_____

CONSIDER $5 **CONDITION**
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.
$_____

Results:

In the control, $5 made up 12% of responses and $10 made up 74%. In the Consider $5 condition, $5 made up 32% and $10 made up 54%. Most respondents still thought the ball cost $10 (95% confidence interval: 51% to 58%), even after being told to consider whether the answer could be $5.

We can use an additional condition to attempt to decompose the 20% increase in $5 response into those who mechanically copied the suggested value into the answer blank and those who solved the problem as a result of $5 being suggested. We administered a “Consider 33” condition to an additional 1,001 participants. It was identical to the Consider 5 condition, except instead of asking participants to consider whether $5 could be the answer, it asked participants to consider whether $33 could be the answer. 8% of those participants said that the ball cost $33 (up from 0% in the control). So, we reason that the consider $5 condition’s 20% increase in $5 response can be decomposed into at least an 8% increase as a result of participants copying that value into the answer blank and as much as a 12% increase as a result of participants actually solving the problem because of our suggestion.

Discussion:

We were surprised that the majority of participants rejected the correct answer in favor of the intuitive error.

In the next series of experiments, we manipulated the intrinsic plausibility of the intuitive error. When the values in the problem are $1.10 and $1.00, their difference ($0.10) is a plausible price for the cheaper of two items whose prices sum to $1.10. We predicted that solution rates would rise if those values were altered so that their difference yielded a less plausible candidate for price of the ball.

7. Plausibility of the Intuitive Error

Frederick (2005) reported that respondents did markedly better on a variant of the bat and ball problem, in which a banana and a bagel cost 37 cents in total, with the banana costing 13 cents more than the bagel. He proposed that performance on the Bat & Ball problem could be dramatically increased by reducing the fluency of the subtraction operation that yielded the erroneous 10 cent response.

Though we replicate Frederick's banana and bagel result,[27] we reject his account of it. Although we presume that people do subtract 13 from 37 somewhat more slowly than they subtract $1.00 from $1.10, we doubt that this explains the elevated performance on this variant. Instead, we suggest that the higher solution rates arise because subtracting the smaller value from the larger one yields a result that is not merely incorrect, but *intuitively* nonsensical. Specifically, if the two items cost $0.37 in total with the cheaper one costing 13 cents less, subtracting the smaller value yields a case in which the *cheaper* item would account for *most* of that total (24 of the 37 cents). By contrast, in the standard problem, subtraction yields a response ($0.10) that easily passes a cursory plausibility check: it is much less than half of the total, and, thus, a possible price for the cheaper of the two items.

[27] Using 2,000 Google survey respondents and bat and ball as the two items to eliminate the slight (but unnecessary) confound with names of the objects, we replicated Frederick's (2005) "banana and bagel" finding; solution rate in the permuted version was higher than in the original (35% vs. 19%; $z = 8.16$, $p < .001$).

Studies 7a-g (Manipulating the difference in cost between the bat and ball)

We conducted seven experiments with a total of 8,280 participants from five different populations.[28] In all studies and all conditions, the first sentence of the problem was identical: “A bat and a ball cost $1.10, in total.” We manipulated the specified difference between the titular items prices: The bat costs [X] more than the ball. Each study contained a standard problem condition (in which X = $1.00) and one or more conditions in which X was smaller.

Results:
Shown below, as X decreases, solution rates increase (Overall: $z(7{,}705) = -30.0$, $p < .001$) [29,30]. The percentage making the intuitive error (responding with the difference between the two printed numbers) is reported as a subscript.[31]

% Correct $_{\text{\% Intuitive Error}}$

Condition:	**A. eLab** *N* = 304	**B. Yale** *N = 41*	**C. Google**[32] *N* = 3,945	**D. mTurk** *N* = 560	**E. Boston** *N* = 534	**F. mTurk** *N* = 321	**G. Google** *N* = 2,008
X = $1.00	29_{66}	48_{33}	19	28_{62}	32_{34}	32_{53}	12_{76}
$0.88	45_{39}	45_{30}	26	26_{60}	--	--	--
$0.60	--	--	--	--	38_{24}	--	--
$0.50	--	--	--	--	45_{12}	--	--
$0.22	--	--	--	54_{21}	--	--	--
$0.12	--	--	57	--	--	--	--
$0.10	--	--	64	63_{21}	56_{6}	57_{15}	46_{30}

Overall, solution rates went from 19% when the bat was $1.00 more than the ball to 56% when it was $0.10 more than the ball. These results both replicate Frederick’s (2005) “banana and bagel” effect and undermine his explanation of it. Whereas he conjectured that difficulty of subtraction increased performance on the banana and bagel problem, our respondents did much better when the difference was $0.10 than when it was $0.88, though the latter subtraction was surely more difficult.

[28] Our analysis omits 567 participants who reported having seen the problem before.

[29] Here and throughout, a failure to respond is counted as an incorrect response.

[30] Includes a matrix of dummies for experiment. As is obvious from the table, results replicate without those controls as well. But in terms of AIC, these solution rates can be most parsimoniously modeled by the single linear effect of the difference and a matrix of dummies for experiment.

[31] As part of his (2012) doctoral dissertation, Jarbas Silva found that just 8% of Brazilian students correctly answered the standard bat & ball problem, whereas 93% could do so when their total cost was 3 cents, with the bat costing 1 cent more.

[32] Here, we omit the percentage making the intuitive error because these data come from binary choices between the correct answer and the intuitive error. The percentage choosing the intuitive error is the compliment of the percentage choosing the correct answer.

Discussion

Like Frederick's Banana and Bagel problem, these data serve to show that many of the people who succumb to the intuitive error are actually capable of the required math. But unlike Frederick's problem, these data make it clear that the crucial factor is some kind of similarity between the intuitive error and the correct answer.

The reader might be tempted to describe these data as showing that the problem becomes easier as the price difference between the bat and ball diminishes.[33] But that description falls short in a number of ways. The table below[34] shows that as the difference shrunk, participants took longer to respond ($t(1177) = -5.44$, $p < .001$).[35] Further, despite their increasing accuracy, their confidence in their answers did not increase.[36,37] In fact, they were more confident with the $1.00 difference than with the $0.10 difference (80% vs. 73% reporting a 100% chance of being correct, $z(567) = 1.95$, $p = .051$; mean confidence 95 vs. 91%, $t(567) = 2.50$, $p = .013$).

Response Time in Seconds (Geometric Means)

Condition:	**A. eLab** *N = 304*	**B. Yale** *N = 41*	**C. Google** *N = 3,945*	**D. mTurk** *N = 560*	**F. mTurk** *N = 321*	**G. Google** *N = 2,008*
X = $1.00	23	14	14	32	19	16
$0.88	45	30	20	56	--	--
$0.22	--	--	--	96	--	--
$0.12	--	--	18	--	--	--
$0.10	--	--	18	54	34	21

When the difference between the printed numbers is a small fraction of the total cost participants offer that difference as their response. But when the difference between the printed numbers is no longer in the ballpark of the ball's price, they reject that error and persist to calculate the ball's true cost. We suspect that this is a feature of heuristic response more generally. People will only rely on the output of a simplifying heuristic when that output is consistent with some rough generalization of the target judgment's requirements. Interestingly, the problem details that inform that rough generalization do not necessarily figure into the heuristic process itself. Here, the fact that the ball price ought be a small fraction of the total is totally extrinsic to the heuristic subtraction. Yet the degree to which it holds still affects endorsement of that heuristic.

[33] It's misleading to presume a linear relation between difference and any measure of difficulty. Imagine the two extremes: At one extreme, the difference is $1.10 (the ball is free) and at the other the difference is zero (the two items cost the same). We suppose that each of these would be an easier problem, in every sense, than any intermediate difference.

[34] The table omits response time data from experiment e. (Boston) because it was paper and pencil.

[35] Google did not give us individual response times, just medians, which we report in the table. So, the reported t statistic just describes mTurk and eLab data. But the Google data are obviously consistent with that result.

[36] Participants in the two mTurk experiments reported their confidence in their answer on an 11 point probability scale from 0% to 100% chance of being correct.

[37] They tended to be less confident as the difference shrunk, p = .11.

General Discussion

The literature on human reasoning includes a number of classic problems, like the Wason Selection task (Wason, 1968) and the Linda problem (). The cognitive psychology literature reports extensively on the Stroop Color-word Effect (Stroop, 1932). And the literature on visual perception includes a variety of experiments on the Muller-Lyer () and Ponzo () illusions. These “fruit flies” of the psychology literature allow scientists to test various explanations of human behavior.

In this paper, we propose a weaker version of the attribute substitution hypothesis to describe why people think the two objects in the bat and ball problem cost $1.00 and 10 cents. Under it, people begin to solve the wrong problem, but note the disparity. However, despite noting the disparity between the actual problem and the one that inspired their response, they fail to realize that their responses violate the actual problem’s constraints. We present evidence for a novel inhibitory process to explain that failure.

We also test three boundaries of the illusion. We find that it largely evaporates when the hypothesized substitute no longer suggests a plausible candidate response for the actual question, but largely persists in the face of manipulations meant to decrease confidence in intuition, and increase salience of the correct answer.

We suspect that our explanation of this particular error is not unique. We hope that its analysis will be applicable to the study of intuition more generally. For example, the inhibitory process that we document is almost maximally relevant to Sloman’s (1996) simultaneity criterion for the existence of separate systems of thought. Specifically, its existence might prevent simultaneous contradictory belief. On the other hand, this inhibition might be specifically evolved to prevent such contradictions, which would only be the case if in its absence they were both common and detrimental.

Appendix A

…

Appendix B: Study B (Bolding "more than the ball")

We used Google Surveys[38] to randomly assign 2,504 participants to either solve the standard bat & ball problem or to solve a variant in which the words, "more than" were bolded. In this experiment, the response format was a binary choice between 5 cents and 10 cents, rather than the standard open-ended response.

CONTROL
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? 5 cents OR 10 cents

***BOLD* CONDITION**
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? 5 cents OR 10 cents

Results:

Replicating experiment 2, there was no significant effect of the bolding manipulation. 5 cent response merely increased from 20% to 23%.

[38] Google pays content providers to post our questions on their webpages. Whenever internet browsers arrive at one of those webpages, they must answer our question before they can see the page. Although participants are randomly assigned to condition, there are selection issues because many participants choose not to respond after seeing the manipulation. In no case are we aware of a selection-based account that contradicts our own. But we are conscious of their possibility.

Appendix C: Studies C1 & C2 (Removing "The bat costs $1.00")

In the following two experiments, we manipulated the problem's wording so that the posited variant could not be created by simply omitting the final four words. We used Google surveys to randomly assign 2,005 web browsers either to solve the standard bat and ball problem or solve a variant in which "The bat costs $100 more than the ball" is replaced by "Their prices differ by $100. The ball is cheaper."

Study C1 (N= 2,005; Google Surveys)

CONTROL
A bat and a ball cost $110 in total. The bat costs $100 more than the ball.
How much does the ball cost? $______

DIFFER
A bat and a ball cost $110 in total. Their prices differ by $100. The ball is cheaper.
How much does the ball cost? $______

Study C2 (N= 770; eLab[39])

CONTROL
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? $______

LESS THAN
A bat and a ball cost $1.10 in total. The ball costs $1.00 less than the bat.
How much does the ball cost? $______

We randomly assigned 770 eLab participants either to solve the standard bat and ball problem or to solve a variant in which the "The bat costs $1.00 more than the ball" is replaced by "The ball costs $1.00 less than the bat."

Results

The table below displays the solution rates and percentages making the intuitive error (as a subscript).

% 5 cents $_{\%\ 10\ \text{cents}}$

Condition:	**C. Google** $_{N\ =\ 2{,}005}$	**D. eLab** $_{N\ =\ 607}$
Control	8_{78}	43_{50}
Differ	10_{73}	--
Less	--	31_{56}

Once again, we were surprised to find that neither manipulation increased performance. In fact, the unexpected performance *decrease* in experiment D was probably not a coincidence ($z(605)$ = 2.XX, p = .013).

[39] We exclude 163 participants who reported having seen the problem before.