Acta Psychologica 164 (2016) 56–64

Contents lists available at ScienceDirect

Acta Psychologica

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

The Doubting System 1: Evidence for automatic substitution sensitivity

Eric D. Johnson [a,*], Elisabet Tubau [a], Wim De Neys [b]

[a] Department of Basic Psychology and IR3C, University of Barcelona, Barcelona, Spain
[b] CNRS, LaPsyDE (CNRS Unit 8042), Paris Descartes University, Paris, France

ARTICLE INFO

Article history:
Received 19 May 2015
Received in revised form 2 December 2015
Accepted 14 December 2015
Available online xxxx

Keywords:
Cognitive reflection
Bat-and-ball problem
Dual process
Executive resources
Reasoning
Decision making
Bias

ABSTRACT

A long prevailing view of human reasoning suggests severe limits on our ability to adhere to simple logical or mathematical prescriptions. A key position assumes these failures arise from insufficient monitoring of rapidly produced intuitions. These faulty intuitions are thought to arise from a proposed substitution process, by which reasoners unknowingly interpret more difficult questions as easier ones. Recent work, however, suggests that reasoners are not blind to this substitution process, but in fact detect that their erroneous responses are not warranted. Using the popular bat-and-ball problem, we investigated whether this substitution sensitivity arises out of an automatic System 1 process or whether it depends on the operation of an executive resource demanding System 2 process. Results showed that accuracy on the bat-and-ball problem clearly declined under cognitive load. However, both reduced response confidence and increased response latencies indicated that biased reasoners remained sensitive to their faulty responses under load. Results suggest that a crucial substitution monitoring process is not only successfully engaged, but that it automatically operates as an autonomous System 1 process. By signaling its doubt along with a biased intuition, it appears System 1 is "smarter" than traditionally assumed.

1. Introduction

In the face of difficulty, human reasoners often appear to forego the effortful processing that may be required and opt instead for less demanding intuitive responses (Kahneman, 2011). While many fast and frugal heuristics are no doubt adaptive in complex and reoccurring environments (Gigerenzer, 2007), thinking fast can also lead to quite embarrassingly erroneous responses in less routine settings. Quickly consider the following example:

A bat and a ball together cost $1.10.
The bat costs $1 more than the ball.
How much does the ball cost?

Intuitively, the answer "10 cents" quickly springs to mind. In fact, a majority of university students, including those from elite schools such as MIT and Harvard, respond with this *intuitive*—but *incorrect*—answer (e.g. Bourgeois-Gironde & Van der Henst, 2009; Frederick, 2005). If a bat costs $1 more than a 10-cent ball, the bat itself must cost $1.10. Summing up, a $1.10 bat + a $0.10 ball would equal $1.20, not $1.10 as stated in the problem. Does this imply that highly educated young adults think that '110 + 10' = '110'? Of course not. Rather, it suggests that even educated reasoners often do not invest the necessary effort needed to correct their initial intuitions, and instead settle for a quickly derived response.

This resistance to cognitive expenditure, or "miserly" thinking, has been most famously characterized by Kahneman (2011), Kahneman and Frederick (2002). According to this dual-process view, when people are confronted with a difficult question, an autonomous System 1 quickly and unconsciously substitutes an easier question in its place. In the bat-and-ball problem, this presumably involves the swapping of the critical relational "more than" statement with a simpler absolute interpretation. That is, people will read "the bat costs more than" as simply "the bat costs", and therefore perhaps ironically give the right answer to the wrong question. Correcting this faulty intuition is assumed to depend on the active monitoring of System 1 by a deliberate and resource-demanding System 2. Due to the human tendency toward miserly or "lazy" thinking, however, this monitoring process typically fails to engage. Without the engagement of System 2, we blindly go with the substituted System 1 response.

More recent work, however, has questioned the extent to which this substitution process goes unnoticed. De Neys, Rossi, and Houdé (2013) solicited participants' judgments of confidence in their response after solving the standard bat-and-ball problem or the following control version:

A magazine and a banana together cost $2.90.
The magazine costs $2.
How much does the banana cost?

In this control version, people will tend to parse the $2.90 into $2 and 90 cents just as naturally as they parse $1.10 in the standard version. However, the control version no longer contains the relative statement

* Corresponding author at: Departament de Psicologia Bàsica, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron, 171, 08035 Barcelona, Spain.
E-mail address: eric.johnson@ub.edu (E.D. Johnson).

http://dx.doi.org/10.1016/j.actpsy.2015.12.008
0001-6918/

("$2 more than the banana") which triggers the substitution. That is, the control version directly presents the easier statement that participants are supposed to be unconsciously substituting in the standard version. If participants are completely unaware that they are substituting when solving the standard version, the standard and control version should be isomorphic and response confidence should not differ. De Neys et al. (2013) observed, however, that participants were much less confident when they erroneously substituted the "10 cents" response on the standard bat-and-ball problem compared to their confidence on the control version (see also Gangemi, Bourgeois-Gironde, & Mancini, 2015, for similar findings). This work suggests that, at least at some level, we are not blind to the substitution process—even biased reasoners showed elementary substitution sensitivity. If this is true, however, it raises an even more fundamental question regarding the source of this sensitivity.

In the present study we contrast two possible origins of this previously observed substitution sensitivity. First, this detection process may be part of a monitoring component of System 2, as suggested by Kahneman (2011), Kahneman and Frederick (2002). On this view, although a supervisory System 2 may not be allocating sufficient resources to the override processes needed to solve the bat-and-ball problem, it is to some extent monitoring for inappropriate output. Bluntly put, System 2 would be more active than typically assumed. However, a second possibility is that this substitution sensitivity arises out of an autonomous System 1 process. On this account, System 1 does not ignorantly throw out an answer whose outcome is at the complete mercy of a vigilant, interventionist System 2. Rather, it sends with its rapid approximation a signal of doubt. Simply put, while System 1 may not be "intelligent" in the traditional sense, neither is it as "dumb" or blind as characteristically assumed.

These two possibilities can be teased apart using the basic processing assumptions of dual process theories. System 1 processes are thought to operate automatically, out of the grip of more controlled, demanding System 2 processing which depends on the availability of executive resources (Evans, 2008; Evans & Stanovich, 2013). The locus of substitution sensitivity can therefore be tested by experimentally manipulating the executive load placed on participants as they reason with the bat-and-ball problem. If detecting an erroneous substitution process is in the domain of a deliberate System 2, then under a resource-demanding load reasoners should not detect this substitution, or this sensitivity should be greatly reduced. If, on the other hand, substitution sensitivity is the work of an automatic System 1 process, then this detection mechanism should be unaffected by load.

In the present investigation we probe this substitution sensitivity in the bat-and-ball problem (and a control version) while reasoning under cognitive load. Four load conditions were used—no load, low load, high load, and extra-high load—to examine the relative contributions of executive resources both for correctly solving the problem and for detecting the presumed substitution when answering with an erroneously substituted response.

In order to validate these findings, we included three different substitution sensitivity measures: Confidence judgments, confidence latencies, and reasoning latencies. Note that the sensitivity findings of De Neys et al. (2013) were based purely on a confidence measure. However, studies investigating basic cognitive control processes in reasoning have shown that decision uncertainty associated with conflict also affects response latencies (Scherbaum, Dshemuchadse, Fischer, & Goschke, 2010; see also Bonner & Newell, 2010; De Neys & Glumicic, 2008; Mevel et al., 2014; Pennycook, Fugelsang, & Koehler, 2012; Stupple & Ball, 2008; Stupple, Ball, & Ellis, 2013; Thompson, Striemer, Reikoff, Gunter, & Campbell, 2003; Villejoubert, 2009). Therefore, if sensitivity arises out of the substitution process then, in addition to reduced response confidence, we should also expect to see longer response times as reasoners attempt to solve the standard version of the task. That is, if reasoners are questioning whether their substituted response is warranted, this uncertainty should translate into increased processing time on the standard bat-and-ball problem relative to a situation where there is no questioning of the immediate intuition (i.e., the control version). Furthermore, latencies for the confidence judgment itself might be affected. If one feels unsure of their response, it may take more time to translate this feeling into a precise estimate of confidence compared to when one is fully confident. Hence, measuring the time it takes to provide a judgment of confidence may provide an additional index of substitution sensitivity.

In sum, if reasoners are sensitive to the substitution process then one can predict that, in addition to previously observed lower confidence ratings, responding to the problem and providing a subsequent judgment of confidence should take longer for standard versus control versions of the task. The key question, however, is whether or not these three detection measures still indicate substitution sensitivity under cognitive load. If this sensitivity depends on the operation of an executive resource-demanding System 2, then its effectiveness should decline under load. However, if substitution sensitivity arises out of autonomous System 1 processes, these measures should be unaffected by load.

2. Experiment

2.1. Method & material

2.1.1. Participants

A total of 324 undergraduate students from the University of Barcelona were recruited for this task in exchange for course credit. Eleven of these students reported being previously familiar with the bat-and-ball problem, and therefore only data from the remaining 313 participants (266 female, 47 male; mean age = 20.50, $SE = 0.28$) was analyzed and reported here.

2.1.2. Reasoning task

The reasoning tasks included a standard and a control version of the bat-and-ball problem introduced above. As in previous work (De Neys et al., 2013), different contextual and numerical contents were used (see Appendix A). One problem presented a bat and ball, the other presented a magazine and banana. In one problem the total cost was $1.10 with one item costing $1 more than the other; in the other problem the total cost was $2.90 with one item costing $2 more than the other. Item contents and values for the standard and control versions were fully counterbalanced across participants, which helps to ensure that any observed effects are general and not driven by the specific material used (e.g. the ease of partitioning 10 from 1.10, or background beliefs about the price of specific items).[1] A blank box with the label "cents" appeared on screen following the problem. Participants therefore typed only their numerical response into the box.

2.1.3. Confidence measure

Immediately following response to either the standard or control version of the reasoning task, participants were asked to indicate how confident they were that their response was correct. Confidence judgments were indicated with a numerical value between 0% (*not at all confident*) and 100% (*completely confident*). As in previous studies (e.g. De Neys, Cromheeke, & Osman, 2011; De Neys et al., 2013), the interest is in the *relative* difference between confidence judgments on the standard substitution version and the control problem. There are numerous reasons for individual variation in absolute ratings of confidence, and a variety of measurement biases may influence the particular value that participants report (e.g. Berk, 2006; Shynkaruk & Thompson, 2006). Accordingly, absolute confidence levels must be interpreted with caution. At the same time, however, it can be assumed that any general bias in the response scale should affect confidence ratings in both standard and control versions. Observing relatively lower confidence

[1] None of these factors had any impact on performance.

following the standard bat-and-ball problem, in particular for those reasoners who provide the erroneous substituted response, can therefore be taken as an indicator that this substitution process has not gone unnoticed.

2.1.4. Load task: dot memory

In the load conditions, participants were presented a secondary visuospatial storage task (De Neys, 2006a; Franssens & De Neys, 2009; Trémolière, De Neys, & Bonnefon, 2012). Prior to the reasoning task, a pattern of dots was briefly presented in a grid for participants to memorize and keep in mind while reasoning (see Fig. 1 for examples). After the reasoning task, participants were subsequently presented a blank grid into which they clicked with the mouse to reproduce the remembered pattern (an indicated dot could be removed by clicking again). In addition to a no load (NL) condition in which the secondary storage task was not presented, three additional load conditions were used in the present study. In the low load (LL) condition, three dots were presented in a single column of a 3 × 3 grid, which should place only a minimal burden on executive resources (De Neys, 2006a; De Neys & Verschueren, 2006). In the high load (HL) condition, four dots were presented in a complex interspersed pattern in a 3 × 3 grid, which has been established to interfere specifically with effort-demanding executive resources (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). In the extra-high load (EHL) condition, five dots were interspersed in a 4 × 4 grid, which has been effectively used to further increase cognitive demands in previous studies (e.g. Trémolière et al., 2012). The dot patterns were briefly presented prior to the reasoning task. Presentation time for the 3 × 3 grids was set to 900 ms. To make sure that participants could perceive the extra complex pattern in the 4 × 4 grid so that storage would effectively burden executive resources, presentation time was increased to 1600 ms in this extra high load condition (i.e. equaling a 100 ms presentation time for each of the nine or 16 quadrants in the grid; see e.g., Trémolière et al., 2012, for a related approach).

2.1.5. Procedure

All tasks were adapted for computer-based testing to allow the collection of response times. Participants were tested in small groups (up to four at a time) at individual computer terminals. All participants were randomized to receive the standard (n = 158) or the control (n = 155) problem in one of the four load conditions (NL, LL, HL, EHL). Appropriate task instructions were provided, along with a brief practice series to familiarize them with the testing environment, explained as follows.

All participants first saw a simple and unrelated math story problem where they were to provide a single numerical response and press the Enter key. On a new screen participants were then instructed to provide a numerical confidence judgment regarding the correctness of their previous response. In the load conditions, this was followed by another screen with instructions explaining that they would also have to memorize a dot pattern to subsequently reproduce after the reasoning task and confidence judgment. Participants in the load conditions then practiced the entire series: A practice pattern was briefly flashed, followed by the same simple practice problem, followed by response confidence, and finally a blank grid appeared for participants to reproduce the previously seen dot pattern. The practice procedure was the same in the no load group, but without any mention or practice of the dot pattern task.

Following the practice series the actual experiment began. The importance of remembering the dot patterns was emphasized in the load groups. No instructions regarding response speed were mentioned in any group. At the end of the experiment two additional control questions were asked. First, participants were asked a confidence-control question to ensure that they were paying attention to the confidence questions. We presented a blatantly false statement (i.e. "How confident are you that Toulouse is the capital of France?").[2] Responses were given on the same 0–100 scale used for the confidence rating. The average rating was 2.25% ($SE = 0.73\%$), with 94.9% of participants entering 0%. Finally, participants were asked to indicate if they were already familiar with the bat-and-ball problem.

2.2. Results & discussion

We first present results for the secondary load task and the impact of load on bat-and-ball response accuracies. Next, we present results for the three critical substitution sensitivity measures.

2.2.1. Accuracy under load

2.2.1.1. Load task. On average, in the 3-dot low load condition participants correctly indicated 100% ($M = 3.00$, $SD = 0.00$) of the dot locations on the standard problems and 94% ($M = 2.83$, $SD = 0.70$) of the dots on the control versions. In the 4-dot high load condition, 91% ($M = 3.64$, $SD = 0.74$) of the dot locations on the standard problems and 81% ($M = 3.24$, $SD = 1.02$) of the dots on the control versions were correctly indicated. In the 5-dot extra-high load condition 94% ($M = 4.68$, $SD = 0.70$) of the dot locations on the standard problems and 83% ($M = 4.13$, $SD = 1.34$) of the dots on the control versions were correctly indicated. This shows that overall the secondary task was performed properly, as participants were instructed to.

There was no correlation between participants scores on the dot recall and reasoning task in the low load, high load ($r = .088$, $p = .59$), or extra-high load conditions ($r = .084$, $p = .62$), which indicates that there was no overall performance trade-off between these tasks. A performance trade-off would imply that higher accuracies on the reasoning task come at the cost of dot recall accuracy. This could indicate that participants strategically neglected the load task. However, although most participants recalled the dot locations correctly, a small minority of participants made some recall errors. In cases where the dot locations are not recalled correctly, one might argue that we cannot be certain that the load task was efficiently burdening executive resources (i.e. the subject might be neglecting the load task, thereby minimizing the experienced load). A possible lack of a load effect on performance could be then attributed to this possible confound. To sidestep this potential problem completely, all subjects who showed imperfect recall (n = 51) were eliminated from the reported analyses.[3]

2.2.1.2. Bat-and-ball accuracy. As shown in Table 1, response accuracy on the control problems was very high across all load conditions. A non-parametric logistic regression on the control versions with accuracy (correct, incorrect) as the dependent variable and load (NL, LL, HL, EHL) entered as a predictor confirmed that performance was clearly not affected by executive load on these control problems ($\chi^2(1) < 1$, $p > .75$). This establishes that the intuitive, and correct, response in the control version was automatically triggered with minimal involvement of executive resources. In sharp contrast, on the standard versions a clear decline in correct responses was observed with increasing load. A second logistic regression performed on the standard versions revealed that this effect of load on accuracy was significant ($\chi^2(1) = 6.53$, $p = .011$, $e^{\beta} = .47$, 95% C.I. = .26–.84). As expected, this directly establishes that correctly solving the classic bat-and-ball problem draws on executive resources.

2.2.2. Substitution sensitivity measures

We next looked at the three substitution sensitivity measures: Confidence judgments, confidence latencies, and reasoning latencies. Recall that the key comparison here with all three sensitivity measures is between participants who provide the *intuitive but incorrect* "10 cents" substitution response on the standard version, and those who

[2] That Paris, not Toulouse, is the capital of France is common knowledge for university students in Barcelona.

[3] Analyses that included these participants were consistent with the reported results. For completeness, the reader can find an overview of the unfiltered data in Appendix B.

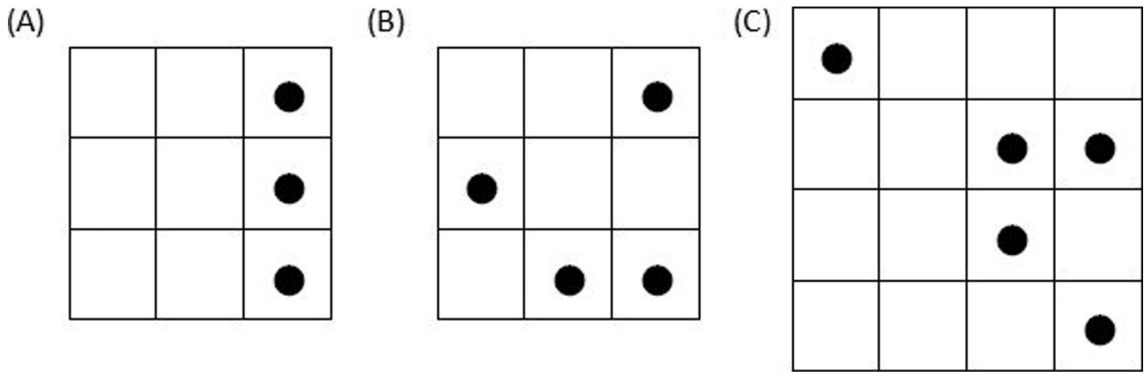

Fig. 1. Example dots patterns presented as a dual-load task in the (A) Low Load, (B) High Load, and (C) Extra-High Load conditions.

give the *intuitive and correct* response on the control version.[4] This enables us to assess whether or not biased reasoners are sensitive to their response substitution. For completeness, we also provide the data for the small group of reasoners who answered the classic version correctly in our Tables. Note that given the very limited number of observations (e.g. only 1 correct response under high and extra-high loads), these latter data should always be interpreted with caution and were not further analyzed.

2.2.2.1. Confidence judgments. An analysis of variance was run on the confidence ratings with version (standard, control) and load (NL, LL, HL, EHL) as independent variables. The analysis revealed a clear main effect of version. As Table 2 indicates, the confidence of participants giving the incorrect "10 cents" substitution response on the standard problem was significantly lower than the confidence of reasoners solving the control version that did not evoke a substitution process ($F(1, 232) = 14.74, p < .001, \eta^2_p = .06$). In line with previous confidence findings (e.g. De Neys et al., 2013), this indicates that biased reasoners were not simply oblivious to their faulty substitution, but indeed detected that their incorrect response was not fully warranted. At the same time, neither the effect of load nor the version × load interaction were significant ($Fs < 1$, $ps > .45$). The lack of significant interaction suggests that this detection process is operating independently of available executive resources.

Although the overall version x load interaction was not significant, visual inspection of Table 2 nevertheless suggests that the confidence decrease tended to be somewhat less pronounced in the high load/extra high load conditions than in the no load/low load conditions. For completeness, we ran a planned contrast collapsing the lowest (no and low load) and highest (high and extra high load) load conditions to test this specific interaction trend directly. However, this test also confirmed that the interaction did not reach significance ($F(1, 236) = 2.2, p > .10$).

2.2.2.2. Confidence response times. Table 3 shows the time it took for participants to provide confidence ratings under varying load when answering with an intuitive response, where that intuition was either correct (control version) or incorrect (standard version). An analysis of variance on log-transformed response times revealed a main effect of version ($F(1, 232) = 9.46, p = .002, \eta^2_p = .04$). Complementing the differential confidence judgments themselves, participants were significantly slower to report their confidence on the standard versions compared to on the control problems.

The effect of load was also significant ($F(3, 232) = 2.95, p = .034, \eta^2_p = .04$), indicating faster responses under load. However, as with the confidence ratings the critical version x load interaction was non-significant ($F < 1, p > .55$). This suggests that substitution sensitivity was independent of load. As with the confidence rating, a further control analysis collapsing the highest and lowest load conditions also failed to reveal an interaction with version on confidence response times ($F(1, 236) < 1$ 2.2, $p > .40$).

2.2.2.3. Reasoning response times. Table 4 shows the time it took for participants to respond to the standard or control versions under varying load. An analysis of variance was run on log-transformed latencies, with version (standard, control) and load (NL, LL, HL, EHL) as independent variables. Consistent with the above indices of substitution sensitivity, a main effect of version ($F(1, 232) = 53.89, p < .001, \eta^2_p = .19$) indicated that, although biased reasoners failed to correct their substituted intuitions, these same individuals were spending significantly more time (on average nearly twice as long) responding than reasoners answering control versions that did not invoke the substitution process. This again suggests that participants were sensitive to the substitution despite their ultimately erroneous response.

A main effect of load ($F(3, 232) = 3.34, p = .020, \eta^2_p = .04$) revealed that participants tended to speed up their response under load. Importantly, however, version did not interact with load ($F < 1, p > .40$). Nevertheless, although the overall version x load interaction was not significant, close inspection of Table 2 might suggest that the latency increase on standard versus control problems tended to be somewhat less pronounced in the high/extra high load conditions compared to the no load/low load conditions. As with the other substitution sensitivity measures, we therefore ran a planned contrast that directly tested this specific interaction trend. However, the effect did not reach significance, ($F(1, 236) = 2.43, p > .10$). This again confirms that the substitution sensitivity is conserved under load.

2.2.2.4. MANOVA results. All three substitution sensitivity measures consistently showed that substitution sensitivity was not affected by load; in no case was there a version x load interaction. Despite this consistency, however, our conclusion is based on acceptance of the

Table 1
Percentages of correct response (standard error) on the control and standard versions under no load (NL), low load (LL), high load (HL), and extra-high load (EHL).

Load	Control version		Standard version	
	% CR	n	% CR	n
NL	97.4 (2.6)	38	21.6 (6.9)	37
LL	100.0 (0)	34	15.9 (5.6)	44
HL	100.0 (0)	25	3.3 (3.3)	30
EHL	95.8 (4.2)	24	3.3 (3.3)	30
Average	98.3 (1.2)	121	12.1 (2.8)	141

[4] Two participants who did not provide the intuitive "10 cents" response on the control problem were removed from the following substitution sensitivity analyses. An additional participant was also discarded for providing a confidence of "100%" on both their bat-and-ball problem as well on the "Toulouse confidence-control question", indicating that this participant was haphazardly responding to the confidence measure. Removal of these subjects did not significantly change results.

Table 2
Confidence judgments (standard error) of participants answering correctly on the control version and participants answering incorrectly (with the substituted "10 cents" response) and correctly on the standard version of the bat-and-ball problem in the no load (NL), low load (LL), high load (HL), and extra-high load (EHL) conditions. Confidence was significantly lower when providing an *intuitive but incorrect* response on the standard bat-and-ball problem compared to when providing an *intuitive and correct* response on the control version.

Load	Control version		Standard version			
	Correct		Incorrect		Correct	
	Conf (SE)	*n*	Conf (SE)	*n*	Conf (SE)	*n*
NL	98.6 (.81)	36	88.9 (3.8)	29	85.0 (6.5)	8
LL	99.7 (.29)	34	88.9 (3.1)	37	97.9 (1.5)	7
HL	100.0 (0)	25	94.5 (3.8)	27	100.0 (0)	1
EHL	97.8 (2.2)	23	94.3 (3.7)	29	50.0 (0)	1
Average	99.1 (.49)	118	91.5 (1.8)	122	89.1 (4.2)	17

null-hypothesis. Accordingly, in order to validate findings, we also ran a 2 (version) × 4 (load) multivariate analysis of variance (MANOVA) with the three sensitivity measures entered as dependent variables. Because the MANOVA simultaneously tests the three substitution sensitivity measures, it can increase the chances of detecting an effect (namely, a possible version × load interaction) that may not appear with an independently run ANOVA (Hill & Lewicki, 2005; Tabachnick & Fidell, 2012; Stevens, 2002). Pillai's trace statistic was used as it is considered the most robust to potential model violations and the most likely to detect an effect if one is indeed present (Tabachnick & Fidell, 2012; Stevens, 2002).

Results confirmed a strong effect of the problem version on substitution sensitivity ($F(3, 230) = 18.58$, $p < .001$, $\eta^2_p = .20$) and a marginal effect of load ($F(9, 696) = 1.70$, $p = .084$, $\eta^2_p = .02$). Crucially, the version × load interaction was non-significant ($F(9, 696) < 1$, $p > .80$). For completeness, we also note that collapsing the no load/low load and high load/extra high load conditions results in the same strong effect of version ($F(3, 234) = 18.51$, $p < .001$, $\eta^2_p = .19$), and a significant effect of load ($F(3, 234) = 1.70$, $p = .038$, $\eta^2_p = .04$). Most importantly, the version x load interaction remained non-significant ($F(3, 234) < 1.2$, $p > .30$). Hence, these data further confirm that cognitive load does not affect substitution sensitivity.

2.2.2.5. Bayes Factor analysis. As noted above, our key conclusions are based on an acceptance of the null hypothesis (i.e. the absence of a version × load interaction). In the Null-Hypothesis Significance Testing (NHST) approach, a statistical inference is always based on the probability of observing a certain difference (D) if the null hypothesis is true (e.g., if p(D|H0) is less than .05, then reject H0). A general limitation of the NHST approach is that it only allows a binary decision to reject or not reject H0 (Campbell & Thompson, 2012). An alternative to NHST that is emerging in psychological research is a Bayesian analysis of posterior probabilities for H0 vs. H1 (e.g. Masson, 2011; Morey, Rouder, Verhagen, & Wagenmakers, 2014; Wagenmakers, 2007).

Table 3
Latencies (standard error) on confidence judgments in the no load (NL), low load (LL), high load (HL), and extra-high load (EHL) conditions. Response times were significantly slower when providing an *intuitive but incorrect* response on the standard bat-and-ball problem compared to when providing an *intuitive and correct* response on the control version.

Load	Control version		Standard version			
	Correct		Incorrect		Correct	
	RT (SE)	*n*	RT (SE)	*n*	RT (SE)	*n*
NL	4153 (636)	36	4454 (709)	29	6772 (1695)	8
LL	2727 (655)	34	4198 (628)	37	3772 (1812)	7
HL	3161 (764)	25	3864 (735)	27	5557 (4794)	1
EHL	2823 (796)	23	4603 (709)	29	13,098 (4794)	1
Average	3216 (358)	118	4280 (348)	122	5837 (1195)	17

Table 4
Response times (standard error) on the bat-and-ball problem in the no load (NL), low load (LL), high load (HL), and extra-high load (EHL) conditions. Biased substitution responders on the standard version were consistently slower than correct reasoners on the control versions, indicating they were detecting the substitution process in the problem despite their erroneous response.

Load	Control version		Standard version			
	Correct		Incorrect		Correct	
	RT (SE)	*n*	RT (SE)	*n*	RT (SE)	*n*
NL	16591 (2232)	36	30757 (2487)	29	82219 (34359)	8
LL	14056 (2297)	34	29137 (2202)	37	106678 (36731)	7
HL	14184 (2679)	25	21500 (2578)	27	31713 (97181)	1
EHL	15190 (2793)	23	22926 (2487)	29	198514 (97181)	1
Average	15005 (1256)	118	26080 (1221)	122	96160 (22665)	17

Calculation of the posterior probability for H0 is based on the Bayes Factor (BF), which is the odds ratio P(D|H0)/P(D|H1). To validate our findings, we used the MorePower (Campbell & Thompson, 2012) software package that allows computation of the Bayes Factor based on ANOVA results. We computed the Bayes Factor of the Version x Load interaction term for each of our 3 substitution detection measures. Results indicated that the Bayes Factor for each of the three interaction terms was very high (confidence; BF = 1057, P(H0|D) > .99, P(H1|D) < .001; confidence RT: BF = 1300, P(H0|D) > .99, P(H1|D) < .001; reasoning RT: BF = 859, P(H0|D) > .99, P(H1|D) < .002). For completeness, we also calculated the Bayes Factor for the interaction contrast between version and the combined no load/low load and high load/extra high load conditions. Results showed that the Bayes Factor remained substantial (confidence; BF = 5, P(H0|D) > .83, P(H1|D) < .17; confidence RT: BF = 11, P(H0|D) > .92, P(H1|D) < .08; reasoning RT: BF = 5, P(H0|D) > .82, P(H1|D) < .18). Wetzels et al. (2011) have presented a graduated evidence scale for interpretation of the Bayes Factor ranging from anecdotal (BF = 1–3), over substantial (BF 3–10), strong (BF 10–30), very strong (BF = 30–100), to decisive evidence for H0 (BF > 100). Hence, based on Wetzels et al. classification, our data can be interpreted as substantial to decisive evidence for H0 (see also Jeffreys, 1961). Taken together, the results of the Bayesian Analysis further support the conclusion that cognitive load is not affecting substitution sensitivity.

3. General discussion

In the present study, three different measures of substitution sensitivity indicated that reasoners detected that their substituted intuitive answers were not fully warranted. Reasoners providing an intuitive but incorrectly substituted response on the standard bat-and-ball problem were slower to respond, had less confidence in their erroneous response, and were slower to indicate their reduced confidence, compared to reasoners answering control versions where intuition cued the correct response. This confirms and extends previous work showing that reasoners are not completely oblivious to their erroneous responding (De Neys et al., 2013).

Crucially, all three measures indicated that the substitution sensitivity takes place even in the presence of a demanding secondary task load. This establishes the automatic nature of this sensitivity, and implies that a critical substitution monitoring process is active without the involvement of executive resources. In contrast to the System 2 monitoring hypothesis of Kahneman (2011), our results suggest that substitution sensitivity arises out of an automatic System 1 process operating outside the demands of executive working memory. That is, along with a biased intuition, System 1 also seems to signal the questionability of this rapidly produced response.

This raises the additional question regarding the nature of this "substitution sensitivity" which accompanies the erroneous intuition in the bat-and-ball problem. One possibility is that it is akin to the automatic conflict detection process observed in earlier studies (e.g., De

Neys, 2012, 2014). Previous work with a number of classical judgment and reasoning tasks has indeed established that, even when responding erroneously, a conflict detection process is actively signaling that prepotent intuitions are violating logical or probabilistic norms (e.g., De Neys & Glumicic, 2008; De Neys et al., 2011; Stupple, Ball, Evans, & Kamal-Smith, 2011; Thompson & Johnson, 2014), and that this detection process operates effortlessly (Franssens & De Neys, 2009). Sensing that an intuitive answer is not fully warranted is thought to arise from a conflict between competing task cues automatically activated by a problem (see De Neys, 2012, 2014; De Neys & Bonnefon, 2013; Handley & Trippas, 2015; Pennycook & Thompson, 2012; Pennycook, Trippas, Handley, & Thompson, 2014; Thompson & Johnson, 2014; Thompson & Morsanyi, 2012; Villejoubert, 2009). In classic reasoning tasks, these activated cues might involve prior world knowledge or stereotypical beliefs, on the one hand, and learned "logical intuitions" (De Neys, 2012, 2014; Villejoubert, 2009), such as an awareness of the importance of base-rates or an elementary sense of the conjunction rule, on the other.

So what might be the internally conflicting cues arising out of System 1 which lead to the presently observed substitution sensitivity? We hypothesize that this sensitivity is tied to a semantic awareness of the relational term "more than". It is generally accepted that the rapid "10 cents" answer is derived from the ease by which the total '1.10' cost is segmented into '1' and '.10'. At the same time, however, when processing language we automatically interpret meanings of the component parts of the speech (e.g. Carpenter, Miyake & Just, 1995; Ferreira, Bailey, & Ferraro, 2002; Postma, 2000; Sanford & Sturt, 2002; Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982). The idea is therefore simply that while the segmentation process provides a quick heuristic response, automatic linguistic operations will also signal that we have neglected one of those relevant parts when we initially opt for the intuitive "10 cents" response. That is, we detect that we have not fully complied with the semantic meaning of the relational terms read in the sentence. Note that to the extent that efficient language processing develops over years of practice, it might accordingly be expected that younger children with less automated language processing skills would be less likely to detect the conflict in the bat-and-ball problem. Interestingly, recent work has indeed shown that this is the case (Rossi, Cassotti, Agogué, & De Neys, 2013).

To be clear, our suggestion is that the "conflict" that is being detected in case of the bat-and-ball problem arises out of the substitution process itself, and is not, for example, a conflict between the intuitive but erroneous "10 cents" response and the correctly calculated "5 cents" answer. As evidenced by our response accuracies under load, calculating this latter answer depends on deliberate reasoning processes which take substantial effort to complete, making it highly unlikely that the actual correct answer could factor into the observed substitution sensitivity. Rather, it is the act of substituting—of utilizing an easier strategy to answer a more difficult problem—that automatically triggers a signal alerting us to this act. Put differently, our findings do not entail that the response to the difficult question is automatically computed by System 1. That is, knowing that the substituted "10 cents" response is questionable does not entail knowledge of the correct "5 cents" answer.

It is important to stress that although we may automatically detect our erroneous substitutions, this does not imply that we are necessarily good reasoners. Very few people were able to use this detection signal to correct their faulty intuitions. In the absence of load, and in line with several previous studies (e.g. De Neys et al., 2013; Frederick, 2005; Bourgeois-Gironde & Van der Henst, 2009), only roughly 20% of educated reasoners were able to overcome their substitution bias. This substitution detection without successful reflection supports the general belief that miserly processing underpins biased responding (Evans, 2010; Frederick, 2005; Kahneman, 2011; Stanovich, 2010). Our point is simply that, even though a majority of people respond incorrectly, a signal of doubt accompanies this bias (see also Thompson, 2009; Thompson & Morsanyi, 2012). Failure therefore results not from lazy monitoring of shallow outputs, but from a failure to convert this signal into the processing necessary to complete the override. Put differently, people do not fail to detect that they need to think harder, they fail to complete the effortful, hard thinking.

Although our findings were consistent across all three of our adopted measures of substitution sensitivity, one might point to some potential caveats. For example, one might note that despite their weak statistical nature, there nevertheless appeared to be trends toward a slightly hampered substitution sensitivity under the highest levels of load. A critic may therefore suggest that although the substitution process might not be very demanding, it is not completely automatic either, which in turn implicates some very minimal involvement of cognitive resources. In other words, the present load simply might not have been sufficiently high to knock out System 2 completely.

In theory this argument has merit and should not be disregarded offhand. Automaticity claims in a dual task study are always relative to the amount of imposed load. Dual task studies do not allow for categorical claims about the absolute redundancy of executive resources.[5] We can only infer that the process of interest is sufficiently automatic to operate under the imposed load. At the same time, it should be clear that when this argument is pushed to the extreme, it becomes unfalsifiable and essentially vacuous. No matter how demanding the secondary task becomes, any absence of load effect on the process of interest can always be explained away by arguing post hoc that the process of interest requires an even smaller amount of executive resources. To advance our knowledge, the argument needs to be considered within practical limits. With these considerations in mind, the present study provides good evidence for the automaticity of substitution sensitivity. To our knowledge, our most extreme load condition—the 4×4 grid with 5 dot recall—is one of the most complex tasks that has been used in the dual process literature to date (e.g. De Neys, 2006a, 2006b; Trémolière et al., 2012). The load task did result in the predicted decrease on reasoning accuracy. Performance was virtually floored under these high load conditions, indicating that System 2 was sufficiently hampered to render computation of the correct response essentially impossible. In contrast, none of our analyses pointed to a reliable impact of load on any of the three substitution sensitivity measures. We believe that the most coherent and parsimonious explanation within the dual process framework is that the observed substitution sensitivity is not the result of System 2 involvement, but instead arises out of an automatically operating System 1 process.

Another potential critique concerns our response latency data because, in order to indicate the relative cost difference between the two items (e.g. "more than the ball"), conflict versions are slightly longer than control versions (i.e. +4 words). It should be noted, however, that the average adult reading time is around 250 words per minute (Landerl, 2001). This would correspond to only approximately 1 s extra to read the standard versus control version, which clearly cannot account for our data.

Lastly, we also note that in the present paper we have applied Kahneman's characterization of substitution bias in heuristics and biases tasks to the bat-and-ball problem (e.g., see also De Neys et al., 2013; Gangemi et al., 2015; Toplak, West, & Stanovich, 2011, for a similar interpretation). However, as one reviewer suggested, one might prefer to characterize substitution differently and limit its use to a narrower range of tasks or situations. Interpreted as such, our findings would point to error or bias sensitivity in general rather than to sensitivity to the substitution process per se. That is, the claim that people are sensitive to substitution is more specific than the claim that people are sensitive to bias (i.e., the substitution is a possible theoretical explanation for the bias). Hence, our claim that the present findings point to

[5] As one reviewer noted, we might relatedly want to conceive executive resources recruitment itself on a continuum rather than as a categorical dichotomy between System 1 and System 2 (e.g., see Osman, 2013). Hence, apart from practical concerns, there might also be good theoretical arguments against a categorical claim.

substitution sensitivity rests on the assumption that the biased responding in the bat-and-ball results from a substitution process.

In what follows we sketch some directions for future studies. First of all, in the present work each participant solved a single problem (i.e., either a standard or control version in one of four load conditions). We opted for this design because it allows the most stringent test of the substitution sensitivity hypothesis, as it can be argued that presenting multiple problems to the same participant may artificially direct attention to the substitution (De Neys et al., 2013; Kahneman, 2000). By opting for a between-subject design in which participants only see a single problem we are able to sidestep this criticism. At the same time, the between-subject design can also give rise to other concerns. For example, although we randomly allocated a large sample of participants to the different conditions, we cannot rule out possible pre-existing differences in cognitive capacity between participants across conditions. A within-subject design could help to reduce any risks of this potential sampling bias. Future studies could also control for this potential confound by directly assessing participants' cognitive capacity with a standardized test of working memory or fluid intelligence.

The present study focused exclusively on the bat-and-ball problem, however it should be noted that this problem also features as one of the items on the popular Cognitive Reflection Test (CRT; Frederick, 2005). The CRT is a very short, 3-item questionnaire designed to measure peoples' ability to suppress impulsive responding, or a tendency toward miserly processing, in a reasoning context. The test shows good correlations with standard cognitive ability tests, quantitative SAT scores, and some typical heuristics and biases (e.g., Frederick, 2005; Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Toplak et al., 2011; Stupple, Gale & Richmond, 2013). In addition to the bat-and-ball problem, the test consists of two other related items on which people will tend to intuitively substitute. Another interesting direction for future work would be to test the generalizability of the present findings with the other items of the CRT.

Recent work on the CRT has attempted to determine its precise psychometric properties, assess its ability to predict performance on traditional heuristics and biases tasks, specify the factors that affect CRT performance, and clarify the construct(s) measured by the CRT (e.g., Campitelli & Gerrans, 2014; Liberali et al., 2012; Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2015; Sinayev & Peters, 2015; Stupple et al., 2013; Toplak et al., 2011). Although the CRT likely measures multiple constructs, one clear finding is that people higher in numeracy (i.e., the ability to understand and use basic numerical and probabilistic concepts) are much more likely to do well on the CRT (e.g., Campitelli & Gerrans, 2014; Liberali et al., 2012; Sinayev & Peters, 2015). Future studies might therefore look to assess whether numeracy or other thinking dispositions affect the present substitution sensitivity findings, in addition to employing tests of general cognitive ability as mentioned above. Evaluating a range of potential predictors will help to explain more precisely why people fail to correctly solve the bat-and ball problem, even after relatively automatic processes have detected that their intuitive thinking is not on the right track.

To conclude, the present work demonstrates that biased reasoning on the bat-and-ball problem does not result from lazy monitoring of rapid intuitions. Much to the contrary, detecting our errors appears to occur quite automatically. By signaling its doubt along with a biased intuition, it appears System 1 is smarter than traditionally assumed.

Acknowledgments

This research was supported by grants from the Generalitat de Catalunya (FI-DGR 2011), the Spanish Ministry of Economics and Competitiveness (PSI2012-35703), and the Agence Nationale de la Recherche (ANR-JSH2-0007-01).

Appendix A

A.1. Standard versions

A bat and a ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost? ___ cents.

[correct = .05 cents; intuitive = .10 cents].

A bat and a ball together cost $2.90. The bat costs $2 more than the ball. How much does the ball cost? ___ cents

[correct = .45 cents; intuitive = .90 cents].

A magazine and a banana together cost $1.10. The magazine costs $1 more than the ball. How much does the banana cost? ___ cents

[correct = .05 cents; intuitive = .10 cents].

A magazine and a banana together cost $2.90. The magazine costs $2 more than the ball. How much does the banana cost? ___ cents

[correct = .45 cents; intuitive = .90 cents]

A.2. Control versions

A magazine and a banana together cost $2.90. The magazine costs $2. How much does the banana cost?

[correct & intuitive = .90 cents].

A magazine and a banana together cost $1.10. The magazine costs $1. How much does the banana cost?

[correct & intuitive = .10 cents].

A bat and a ball together cost $2.90. The bat costs $2. How much does the ball cost?

[correct & intuitive = .90 cents].

A bat and a ball together cost $1.10. The bat costs $1. How much does the ball cost?

[correct & intuitive = .10 cents].

Appendix B

In the above reported analyses, participants who failed to recall the load pattern correctly were removed. Here we report the full unconditional analysis which does not remove these failed-recall participants (e.g. it may be argued there recall was impaired due simply to the increased load demands, and not necessarily from a mere neglect of the recall task).[6] In doing so, results were completely consistent with the conditional analysis. Crucially, no significant version × load interactions were observed on confidence judgments ($F < 1.1$, $p > .35$), confidence latencies ($F < 1$, $p > .55$), or reasoning latencies ($F < 1$, $p > .55$). This was also the case in an unconditional MANOVA test ($F < 1$, $p > .80$), and also if collapsing the no load/low load and high load/extra high load conditions ($F < 1.2$, $p > .30$). Results from the Bayes Factor analysis were also highly similar (confidence; BF = 991, $P(H0|D) > .99$, $P(H1|D) < .001$; confidence RT: BF = 1745, $P(H0|D) > .99$, $P(H1|D) < .001$; reasoning RT: BF = 1672, $P(H0|D) > .99$, $P(H1|D) < .001$). This was also the case when calculating the Bayes Factor for the interaction contrast between versions with the combined no load/low load and high load/extra high load conditions (confidence; BF = 4, $P(H0|D) > .79$, $P(H1|D) < .20$; confidence RT: BF = 14, $P(H0|D) > .93$, $P(H1|D) < .07$; reasoning RT: BF = 8, $P(H0|D) > .88$, $P(H1|D) < .12$). For completeness, we include the unfiltered accuracy, confidence, and response latencies below.

[6] Two of these participants also provided a confidence of "100%" on both their bat-and-ball problem as well on the "Toulouse confidence-control question", and were therefore not included in the following analysis.

Table A1
Percentages of correct response (standard error) on the control and standard versions under no load (NL), low load (LL), high load (HL), and extra-high load (EHL).

Load	Control version		Standard version	
	% CR	n	% CR	n
NL	97.4 (2.6)	38	21.6 (6.9)	37
LL	100.0 (0)	36	15.9 (5.6)	44
HL	100.0 (0)	41	2.6 (2.6)	39
EHL	97.5 (2.5)	40	2.6 (2.6)	38
Average	98.7 (.9)	155	10.8 (2.5)	158

Table A2
Confidence judgments (standard error) in the no load (NL), low load (LL), and high load (HL), and extra-high load (EHL) conditions.

Load	Control version		Standard version			
	Correct		Incorrect		Correct	
	Conf (SE)	*n*	Conf (SE)	*n*	Conf (SE)	*n*
NL	98.6 (.81)	36	88.9 (3.8)	29	85.0 (6.5)	8
LL	99.7 (.28)	36	88.9 (3.1)	37	97.9 (1.5)	7
HL	100.0 (0)	40	94.3 (3.1)	36	100.0 (0)	1
EHL	98.7 (1.3)	38	94.7 (2.9)	37	50.0 (0)	1
Average	99.3 (.39)	150	91.9 (1.6)	139	89.1 (4.2)	17

Table A3
Latencies (standard error) on confidence judgments in the no load (NL), low load (LL), and high load (HL), and extra-high load (EHL) conditions.

Load	Control version		Standard version			
	Correct		Incorrect		Correct	
	RT (SE)	*n*	RT (SE)	*n*	RT (SE)	*n*
NL	4153 (599)	36	4454 (668)	29	6772 (1695)	8
LL	2719 (599)	36	4198 (591)	37	3772 (1812)	7
HL	2902 (569)	40	4067 (599)	36	5557 (4794)	1
EHL	2831 (584)	38	4253 (591)	37	13,098 (4794)	1
Average	3151 (294)	150	4243 (307)	139	5837 (1195)	17

Table A4
Latencies (standard error) on the reasoning problem in the no load (NL), low load (LL), and high load (HL), and extra-high load (EHL) conditions.

Load	Control version		Standard version			
	Correct		Incorrect		Correct	
	RT (SE)	*n*	RT (SE)	*n*	RT (SE)	*n*
NL	16591 (2145)	36	30757 (2390)	29	82219 (34359)	8
LL	13702 (2145)	36	29137 (2116)	37	106678 (36731)	7
HL	13904 (2034)	40	22475 (2145)	36	31713 (97181)	1
EHL	14381 (2088)	38	23294 (2116)	37	198514 (97181)	1
Average	14645 (1052)	150	26416 (1097)	139	96160 (22665)	17

References

Berk, R. A. (2006). *Thirteen strategies to measure college teaching: A consumer' s guide to rating scale construction, assessment, and decision making for faculty, administrators, and clinicians.* Sterling: Stylus.

Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory & Cognition*, *38*, 186–196.

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, A. P. Bloisdell, L. Huber, & A. Young (Eds.), *Rational animals, irrational humans* (pp. 235–252). Tokyo, Japan: Keio University Press.

Campbell, J. I. D., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavioral Research*, *44*, 1255–1265.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory and Cognition*, *42*, 434–447.

Carpenter, Miyake, & Just (1995). Language comprehension: Sentence and discourse processing. *Annual Review of Psychology*, *46*, 91–120.

De Neys, W. (2006a). Dual processing in reasoning: two systems but one reasoner. *Psychological Science*, *17*, 428–433.

De Neys, W. (2006b). Automatic-heuristic and executive-analytic processing in reasoning: Chronometric and dual task considerations. *Quarterly Journal of Experimental Psychology*, *59*, 1070–1100.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*, 28–38.

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking and Reasoning*, *20*, 169–187. http://dx.doi.org/10.1080/13546783.2013.854725.

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

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

De Neys, W., & Verschueren, N. (2006). Working memory capacity and a notorious brain teaser: The case of the Monty Hall Dilemma. *Experimental Psychology*, *53*, 123–131.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PloS One*, *6*, e15954.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*, 269–273.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278.

Evans, J. S. B. T. (2010). Intuition and reasoning: a dual process perspective. *Psychological Inquiry*, *21*, 313–326.

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

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15.

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking and Reasoning*, *15*, 105–128.

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

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—In search of a phenomenon. *Thinking and Reasoning*, *21*(4), 383–396.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious.* New York: Viking.

Handley, S. J., & Trippas, D. (2015). Dual processes, knowledge, and structure: A critical evaluation of the default interventionist account of biases in reasoning and judgement. *Psychology of Learning and Motivation*, *62*.

Hill, T., & Lewicki, P. (2005). *Statistics: Methods and applications.* New York: StatSoft, Inc.

Jeffreys, H. (1961). *Theory of probability.* Oxford, U.K.: Oxford University Press.

Kahneman, D. (2000). A psychological point of view: violations of rational rules as a diagnostic of mental processes [commentary on Stanovich and West]. *Behavioral and Brain Sciences*, *23*, 681–683.

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Strauss, Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics & biases: The psychology of intuitive judgment* (pp. 49–81). New York: Cambridge University Press.

Landerl, K. (2001). Word recognition deficits in German: more evidence from a representative sample. *Dyslexia*, *7*(4), 183–197.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, *25*(4), 361–381. http://dx.doi.org/10.1002/bdm.752.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*, 679–690. http://dx.doi.org/10.3758/s13428-010-0049-5.

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2014). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*. http://dx.doi.org/10.1080/20445911.2014.986487.

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

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E. -J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming. *Psychological Science*, *25*, 1289–1290.

Osman, M. (2013). A case study: dDual-process theories of higher cognition—Commentary on Evans & stanovich (2013). *Perspectives on Psychological Science*, *8*(3), 248–252.

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base-rates is routine, relatively effortless and context-dependent. *Psychonomic Bulletin & Review*, *19*, 528–534.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, *124*, 101–106.

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

Postma (2000). Detection of errors during speech production: a review of speech monitoring models. *Cognition*, *77*(2), 97–132.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2015). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*. http://dx.doi.org/10.1002/bdm.1883.

Rossi, S., Cassotti, M., Agogué, M., & De Neys, W. (2013). Development of substitution bias sensitivity: Are adolescents happy fools? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*. (pp. 3321–3326).

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*(9), 382–386.

Scherbaum, S., Dshemuchadse, M., Fischer, R., & Goschke, T. (2010). How decisions evolve: The temporal dynamics of action selection. *Cognition*, *115*, 407–416.

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology*, *14*, 489–537.
Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition*, *34*, 619–632.
Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, *6*, 532.
Stanovich, K. E. (2010). *Rationality and the reflective mind*. New York: Oxford University Press.
Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Erblaum.
Stupple, E. J. N., & Ball, L. J. (2008). Belief-logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking and Reasoning*, *14*, 168–181.
Stupple, E. J. N., Ball, L. J., Evans, J. S. B. T., & Kamal-Smith, E. (2011). When logic and belief collide: Individual differences in reasoning times support a selective processing model. *Journal of Cognitive Psychotherapy*, *23*(8), 931–941.
Stupple, E. J. N., Ball, L. J., & Ellis, D. (2013a). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking and Reasoning*, *19*(1), 54–77.
Stupple, E. J. N., Gale, M., & Richmond, C. (2013b). Working memory, cognitive miserliness and logic as predictors of performance on the cognitive reflection test. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1396–1401). Austin, TX: Cognitive Science Society.
Tabachnick, & Fidell (2012). *Using multivariate statistics* (6th ed.). Pearson: New York.
Thompson, V. A. (2009). Dual process theories: A metacognitive perspective. In J. Evans, & K. Frankish (Eds.), *Two minds: dual processes and beyond*. Oxford, UK: Oxford University Press.
Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking and Reasoning*, *20*(2), 215–244.
Thompson, V. A., & Morsanyi, K. (2012). Analytic thinking: do you feel like it? *Mind & Society*, *11*, 93–105.
Thompson, V. A., Striemer, C. L., Reikoff, R., Gunter, R. W., & Campbell, J. I. D. (2003). Syllogistic reasoning time: Disconfirmation disconfirmed. *Psychonomic Bulletin & Review*, *10*, 184–189.
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*, 1275–1289.
Trémolière, B., De Neys, W., & Bonnefon, J. F. (2012). Mortality salience and morality: thinking about death makes people less utilitarian. *Cognition*, *124*, 379–384.
Villejoubert, G. (2009). Are representativeness judgments automatic and rapid? The effect of time pressure on the conjunction fallacy. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*. (pp. 2980–2985).
Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804. http://dx.doi.org/10.3758/BF03194105.
Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. -J. (2011). Statistical evidence in experimental psychology: an empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298. http://dx.doi.org/10.1177/1745691611406923.