Accepted for publication in Cognition. Uncorrected authors' manuscript. January 2016.

The Time Course of Conflict on the Cognitive Reflection Test

Eoin Travers

Queen's University Belfast, UK

Jonathan J Rolison

Queen's University Belfast, UK

Aidan Feeney

Queen's University Belfast, UK

Words (text): 5,905

Address for correspondence:

Eoin Travers

School of Psychology

Queen's University Belfast, UK

David Keir Building

Belfast, BT7 1NN

Email: etravers01@qub.ac.uk

Telephone: +44 (0)28 9097 5653

Accepted for publication in Cognition. Uncorrected authors' manuscript. January 2016.

Abstract

Reasoning that is deliberative and reflective often requires the inhibition of intuitive responses. The Cognitive Reflection Test (CRT) is designed to assess people's ability to suppress incorrect heuristic responses in favor of deliberation. Correct responding on the CRT predicts performance on a range of tasks in which intuitive processes lead to incorrect responses, suggesting indirectly that CRT performance is related to cognitive control. Yet little is known about the cognitive processes underlying performance on the CRT. In the current research, we employed a novel mouse tracking methodology to capture the time-course of reasoning on the CRT. Analysis of mouse cursor trajectories revealed that participants were initially drawn towards the incorrect (i.e., intuitive) option even when the correct (deliberative) option was ultimately chosen. Conversely, participants were not attracted to the correct option when they ultimately chose the incorrect intuitive one. We conclude that intuitive processes are activated automatically on the CRT and must be inhibited in order to respond correctly. When participants responded intuitively, there was no evidence that deliberative reasoning had become engaged.

Keywords: Cognitive Reflection Test; Dual Processes; Reasoning; Conflict; Mouse tracking; Heuristic.

Accepted for publication in Cognition. Uncorrected authors' manuscript. January 2016.

The Time Course of Conflict on the Cognitive Reflection Test

1. Introduction

The Cognitive Reflection Test (CRT; Frederick, 2005) is a brief test designed to measure individuals' ability to inhibit intuitive responses in favor of reflective and deliberative reasoning. In the bat-and-ball problem, one of the best-known CRT items, participants are asked:

"A bat and a ball together cost £1.10.
A bat costs £1 more than a ball.
How much does a ball cost?"

The appealing but incorrect response, to say "10p", is believed to be generated effortlessly and automatically by intuitive processes. Arriving at the correct response of "5p" may require that this intuitive response is inhibited in favor of the result of sustained, effortful deliberation.

The CRT has become a popular measure of individual differences, for example it has been cited 11 times in Cognition since 2012, including 6 experiments using the test. Higher CRT scores predict better performance on various cognitive tasks, including reduced framing effects, less discounting of delayed rewards (Frederick, 2005; Cokely & Kelley, 2009) and probability matching (Koehler & James, 2010), resistance to the illusion of explanatory depth (Fernbach, Rogers, Fox, & Sloman, 2013) and conjunction fallacies (Oechssler, Roider, & Schmitz, 2009), greater metacognitive awareness (Mata, Fiedler, Ferreira, & Almeida, 2013) and less endorsement of supernatural belief (Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012; Shenhav, Rand, & Greene, 2012), as well as performance on various tasks that pit normative responding against intuition (Toplak, West, & Stanovich, 2011). Scores on the CRT correlate with measures of IQ and personality characteristics, and usually predict performance on other tasks even when these are controlled for (Toplak et al., 2011).

The CRT is viewed by some as a prototypical application of dual process theories of cognition (Kahneman & Frederick, 2005; Toplak et al., 2011). Dual process theories (Evans, 2008; Evans & Stanovich, 2013; Kahneman, 2011; Sherman, Gawronski, & Trope, 2014) broadly distinguish Type 1 processes that quickly and effortlessly generate intuitive responses, and Type 2 processes that are under deliberative control and are demanding on working memory resources. Consistent with this, a number of studies (Böckenholt, 2010; Campitelli & Gerrans, 2014; Campitelli & Labollita, 2010) have shown that performance on the CRT is predicted by a combination of dispositional factors, inhibitory control, and numerical ability.

Dual process theories differ in their account of CRT performance. Intuition is the default mode of processing in default-interventionist models (Evans, 2006; Kahneman & Frederick, 2005), which hold that Type 2 processes must be engaged for reflective and deliberative processing to inhibit and override intuitive responses. Failure to engage Type 2 processes has been linked to individual differences in personality and intelligence (Stanovich and West, 2008) and task characteristics (Rolison, Evans, Walsh, & Dennis, 2011). When Type 2 processes are engaged, they may nevertheless fail to adequately replace an intuitive response (Stanovich and West, 2008). Failure to engage Type 2 processes has been proposed to explain incorrect heuristic responses on the CRT. Default-interventionist models make an important prediction about cognitive conflict during reasoning on the CRT. When a heuristic response is given, deliberative Type 2 processing likely has failed to become properly engaged. However, when the correct response is given, the incorrect, Type 1, heuristic response must have been inhibited by Type 2 processing.

In contrast to default-interventionist accounts, parallel-competitive dual process theories (Sloman 1996; 2014) hold that both Type 1 and Type 2 processes are activated simultaneously, and that they compete for control of behavior. In common with default-interventionist models, these accounts predict that Type 1 intuitive responses must be inhibited in order to reason correctly. Uniquely though, parallel models would also predict Type 2 processes should attempt to signal the correct response, even when failing to overrule the output of Type 1 processes.

More recently, De Neys (2012, 2014) has proposed an intuitive logic model. This modifies the traditional default-interventionist model to account for many findings which indicate that when participants provide biased, heuristic responses, they are often implicitly aware of some conflict between their responses and the normative standard. According to this model, Type 1 processes are sensitive to normative principles, such as logical principles in syllogistic reasoning tasks, or mathematical rules on the bat-and-ball problem. As a result, they implicitly signal a conflict when the incorrect heuristic response is given. However, because the heuristic response is usually prepotent, participants often fail to inhibit it, even when they do detect that it conflicts with normative principles. It is unclear at present, however, how this conflict is actually detected. One possibility is that Type 1 processes simultaneously produce both heuristic and correct responses, and it is the conflict between these two partially active responses which is detected directly. Alternatively, the process may be more subtle, with Type 1 processes not generating a fully-formed correct response, but rather detecting, through some other means, that the heuristic response is questionable. Clearly, these two possibilities make different predictions about conflict between competing response options. In the former case, the intuitive logic model would, like a parallel-competitive account, predict that because both responses are partially cued, participants should be drawn towards giving the correct response during reasoning, even when they ultimately give the heuristic one. In the latter case, if Type 1 processes can signal conflict without actually generating the correct response, participants may experience conflict and uncertainty, but not be actually drawn towards the correct response when giving the heuristic one.

Evidence of the implicit conflict detection predicted by the intuitive logic model comes from a range of experimental paradigms (see De Neys, 2012, for a review). Typically, these studies compare conflict problems, in which the intuitive, heuristic response is incorrect, to no-conflict versions, where both heuristics and normative principles cue the same response. Type 1 processes cue both the heuristic response on conflict problems and the correct response on no-conflict problems. If participants detect the conflict between normative principles and their heuristic responses, they should show greater evidence of conflict on these problems, compared to the no-conflict problems. Such conflict has been measured using confidence ratings (De Neys, Cromheeke, & Osman, 2011), response times (De Neys & Glumicic, 2008), neuroimaging (De Neys, Vartanian, & Goel, 2008), and galvanic skin response (De Neys, Moyens, & Vansteenwegen, 2010), among other measures.

Two studies have directly tested the intuitive logic model when applied to the CRT. De Neys, Rossi, & Houdé (2013) showed that heuristic responses on conflict problems were given with less confidence than correct responses on no-conflict problems. Gangemi, Bourgeois-Gironde and Mancini (2014) report similar effects, asking participants to fill out a brief questionnaire measuring their "feeling of error" after answering either the original bat-and-ball problem or a no-conflict control version, both when participants were asked to generate their responses, and when asked to choose between the heuristic and correct responses. These findings all suggest that participants are to some extent aware of the inadequacy of their heuristic responses.

One difficulty in interpreting the above findings is differentiating between *conflict* and *uncertainty*. Conflict requires that participants are drawn towards two responses at the same time — the correct one, and the heuristic one. Uncertainty, on the other hand, does not require that participants are drawn to the correct response when they select the heuristic one, merely that they experience some sense of unease, indecision, or lack of confidence while doing so. It is difficult to say, without additional evidence, whether conflict, or uncertainty, underlie the results of earlier studies of intuitive logic on the CRT.

In this study, we introduce a novel methodology which addresses this issue, and reveals the time-course of cognitive processing during reasoning on the CRT. Participants completed a computer-based multiple-choice version of the CRT while their mouse cursor movements were recorded. Mouse tracking has been used in other areas of psychology to reveal the time course of decisions on the basis of participants' mouse cursor trajectories over a short period of time (Freeman, Dale, & Farmer, 2011; Spivey, Grosjean, & Knoblich, 2005). We employ it here to capture the cognitive processing underlying CRT performance over a longer timescale. If a classic default-interventionist account explains performance on the CRT, participants should exhibit an initial attraction to an incorrect heuristic option when a correct deliberative option is chosen, but not vice versa, when the heuristic option is chosen. If instead a parallel-competitive model explains performance on the task, then participants should also show attraction to the correct option when the intuitive option is chosen. The predictions of the intuitive logic model depend on the nature of the conflict detection process. If participants detect conflict because both responses are simultaneously generated by Type 1 processes, then the intuitive logic model, like the parallel-competitive model, would predict conflict in both directions. Alternatively, if the conflict detection process is more subtle, relying on a feeling of uncertainty, then like the classic default-interventionist account it might predict that participants should be drawn to the heuristic option when selecting the correct one, but not the other way around.

2. Method

2.1 Participants

One hundred and thirty one students at Queen’s University Belfast participated in exchange for course credit.

2.2 Materials

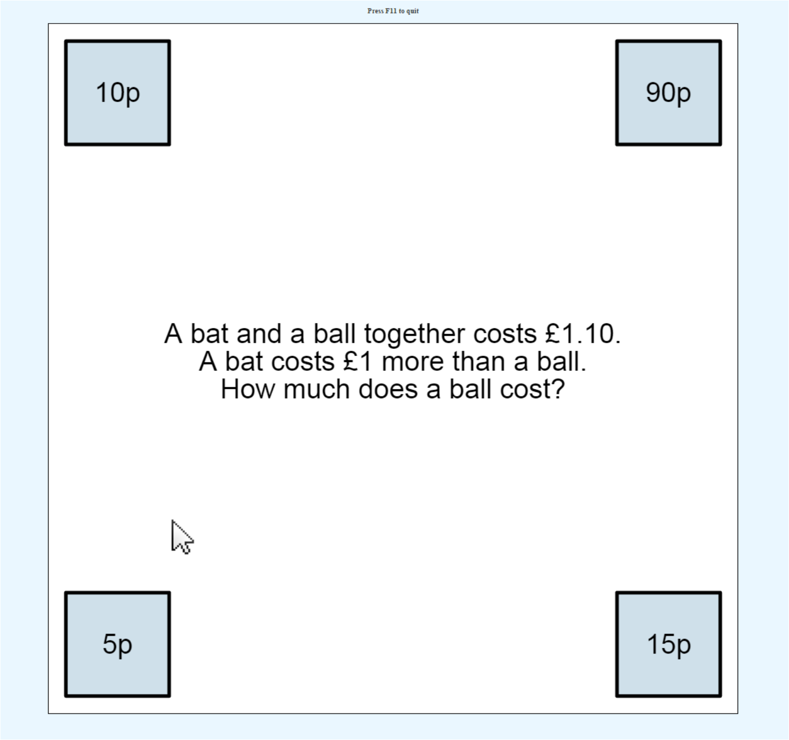
Eight problems were adapted from Primi, Morsanyi, Donati, Chiesi, and Hamilton’s (2015) extended version of the CRT. Each of these problems was modified to create a set of eight corresponding no-conflict problems, in which the intuitively appealing responses were also the correct ones (see the Appendix). Participants were randomly allocated to complete either conflict versions of problems 1, 3, 5, and 7 and no-conflict versions of the rest, or vice versa. Each problem was presented in a 4-option multiple choice format[1]. For the conflict items, the

[1] Although it is unusual to present the CRT as a multiple-choice test, multiple-choice versions have been used previously by Morsanyi, Busdraghi and Primi (2014), Primi et al. (2015; Experiment 3), and Gangemi et al. (2013; Experiment 2), without any clear effect on participants’ responses.

possible responses were the correct option, the incorrect heuristic option, and two incorrect foil options. For the no-conflict problems the correct intuitive option was presented with three incorrect foils.

2.3 Procedure

The experiment was administered on personal computers using custom programmed software. Participants were instructed to respond in their own time to each CRT problem by clicking on one of the four response options presented in the four corners of the display (Figure 1). Participants were not made aware of the mouse tracking aspect of the experiment in advance.



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

Figure 1: An example display from the CRT.

Each item was preceded by onscreen instructions to click on a button marked “Go”, presented in the center of the monitor. This was done to ensure the mouse cursor was located in the same central position at the beginning of each trial. The button was then replaced by the problem text and the four response options appeared simultaneously in the corners (Figure 1). The response options were randomly assigned to the four locations on each trial, with the constraint that the correct and heuristic response options were always adjacent for conflict problems. The mouse cursor was no longer visible at the onset of each trial to prevent it from obscuring the question text. The cursor reappeared once it had been moved more than 5% of the width of the display. Mouse cursor location was recorded every 25 msec.

3. Results

3.1 By-trial analysis

After excluding data from 3 participants who did not complete the experiment within the 15 minutes allocated, and 7 trials with response times greater than 100 seconds (.6% of the total), participants selected the correct option on 80% of no-conflict problems. On the conflict problems, the correct option was chosen 36% of the time, the heuristic option 58%, and one of the foils 6% of the time. A breakdown of responses for each individual problem is shown in the appendix.[2]

[2] Primi et al's (in press) extended version of the CRT was designed to capture more variance than the original three-item version. Problems 5 and 6 are therefore considerably less

In the first stage of our analysis, we calculate a number of summary statistics for each trial, and compare these between problem types, and between responses. The measures were response time, the distance travelled by the mouse cursor (scaled so that a straight line from the start point to the response corresponds to 1 unit), the number of times the cursor was moved during a trial (with movements defined as windows of 100 msec or more in motion, separated by 100 msec or more not moving), and the closest proximity achieved between the cursor and the non-chosen option (closest proximity to the heuristic response option on trials where the correct option was chosen, and vice versa). These measures were compared using linear mixed models, with crossed random intercepts for each participant, and each problem (see Baayen, Davidson, & Bates, 2008). Response latencies, and the distance travelled by the mouse cursor were log-transformed to normalize their distributions, and a generalized mixed model with a Poisson log link was used to model the number of movements. To calculate p values for linear models, degrees of freedom for each parameter were calculated using Satterthwaite's approximation (Kuznetsova, Brockhoff & Christensen, 2015; Satterthwaite, 1946).

difficult than the other problems, and as a result performance on the conflict versions of these problems was similar to that on the no-conflict versions. Therefore, we replicated each analysis reported below on a subset of the data excluding these problems. All reported significant effects were unchanged, or increased in size, and all reported null effects remained, when these problems were excluded.

Consistent with a dual process interpretation, for conflict problems there was greater evidence of conflict across all measures when participants gave the correct response (N = 181) than the heuristic one (N = 297). The average time to respond was 27.3 seconds (SD = 16.3) for correct responses, and 21.0 seconds (SD = 13.4; e^{β} = 1.14, t(470.8) = 2.349, p = .0192) for heuristic responses. The mouse cursor travelled a greater distance before selecting a correct option (6.11 times the minimum needed distance, SD = 5.6) than a heuristic option (5.66 times, SD = 4.74; e^{β} = 1.16, t(298.4) = 2.267, p = .0241). There were also more cursor movements on trials in which the correct response was given (5.4, SD = 4.8) than when the heuristic response was given (4.9, SD = 4.5; e^{β} = 1.15, z = 2.337, p = .0195). Finally, the minimum distance between the cursor and the heuristic option on trials in which the correct option was chosen was on average 49% of the display width (SD = 24%), significantly less than the minimum distance between the cursor and the correct option on trials in which the intuitive option was chosen (55.5%, SD = 18%, e^{β} = 0.92, t(72.1) = 4.119, p < .0001).

Most tests of the intuitive logic model compare correct responses on no-conflict problems with heuristic responses on conflict problems, on the basis that heuristic, Type 1 processes should cue both kinds of response, but the chosen response conflicts with normative principles on conflict problems only. Evidence for the intuitive logic model therefore comes from results which indicate greater conflict for heuristic responses to conflict problems. However, when we calculated each of the applicable measures for correct responses to no-conflict problems (N = 404) we found no evidence of difference between the two types of response: response time (23.1 seconds; t(14.3) = 0.222, p > .8), distance travelled (5.6, SD = 5.0; t(15.0) = 0.359, p > .7) and number of movements per trial (5.2, SD = 4.6; z = 0.064, p > .95).

We also wished to explore any differences in these effects between the various problems. Therefore, we fit an additional mixed model comparing response times for conflict and no-conflict versions of each problem. We included crossed random intercepts for each participant and each problem, and, crucially, we allowed the effect of condition to vary between problems. The full results of this model can be found in the Supplementary Materials. Consistent with the analysis above, there was no significant difference between the conflict and no conflict problems; t(7.4) = .551, p > .5. However, there was some variation between the problems, with the model showing a marginal effect in the direction predicted by the intuitive logic theory for the bat-and-ball problem, a robust effect in the opposite direction for the lily pad problem, and no significant effects for the remaining problems.

Following previous intuitive logic studies (e.g. Mevel et al., 2014; Pennycook, Fugelsang & Koehler, 2015), we also calculated the number of heuristic responses given by each participant on conflict problems, and categorized each of the 128 participants who made at least one heuristic response as either "majority heuristic" (3 or 4 heuristic responses out of four, 53 participants) or "minority heuristic" (1 or 2 heuristic responses, 75 participants). We entered this measure as a participant-level predictor in our models, but found that it was not involved with any interactions in the analyses above (t's < 1, p's > .4). We also repeated this analysis, comparing participants who made the most (four heuristic responses) and fewest (one heuristic response) heuristic responses, and again found no significant interactions (t's < 1.1, p's > .25). Therefore, these analyses revealed no evidence for the existence of conflicting responses regardless of how many heuristic responses participants gave.

3.2 Time course Analyses

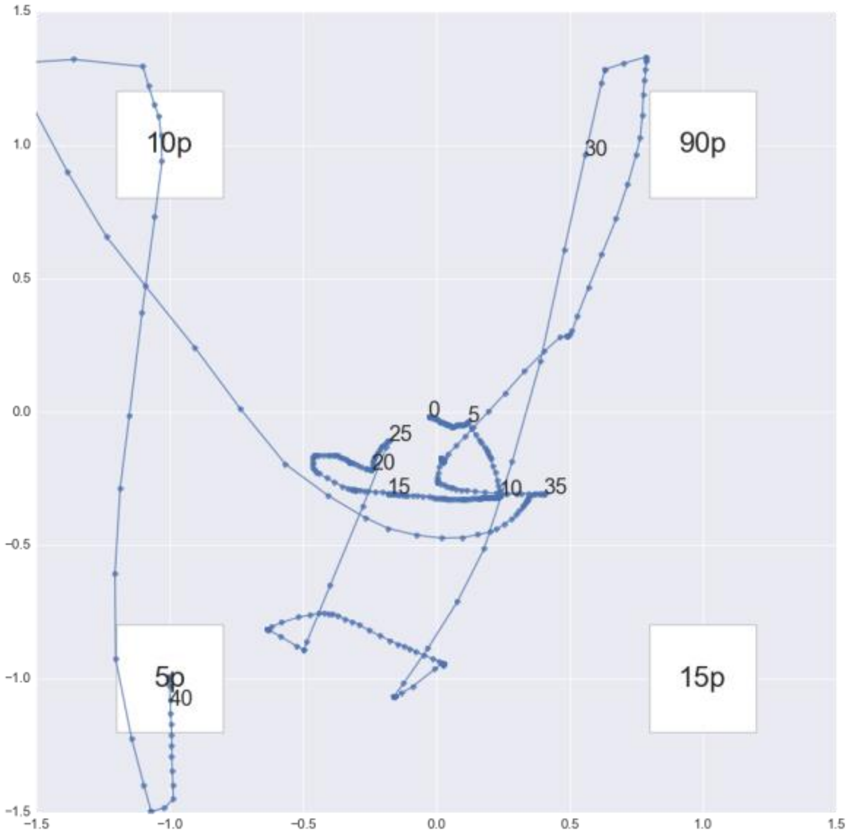


Figure 2: A typical mouse cursor trajectory from the conflict condition. Numerical values indicate the time elapsed in seconds. Cursors meandered as participants generated their responses, passing near the response options located in the corners of the display.

In previous mouse tracking research (e.g. Freeman et al., 2011; Spivey et al., 2005), recording movements over a few seconds, the cursor typically moves straight to a response option, curves between two of them, or in some cases moves to one, and then the other. In our data, unfolding over up to 60 seconds, participants move and rest the cursor many times throughout a trial (an average of 5.1 times, and a maximum of 30), in a manner more similar to that of eye movements. A typical mouse cursor trajectory, shown in Figure 2, comprises a number of movements which pass near to several response options. In order to analyze participants' attraction to each response option over time, the display was divided into quadrants corresponding to the correct option, the intuitive option, and the two foil options. For the first 60 seconds of each trial, the mouse cursor positions at each 200 millisecond time slice were coded according to which section of the screen they occupied, similar to fixation analyses of eye-tracking data. For all the time course plots that follow, additional figures are included in the Supplementary Materials showing these data plotted separately for each problem, and for "majority/minority heuristic" participants. However, the results appear to be broadly consistent across problems and participants.

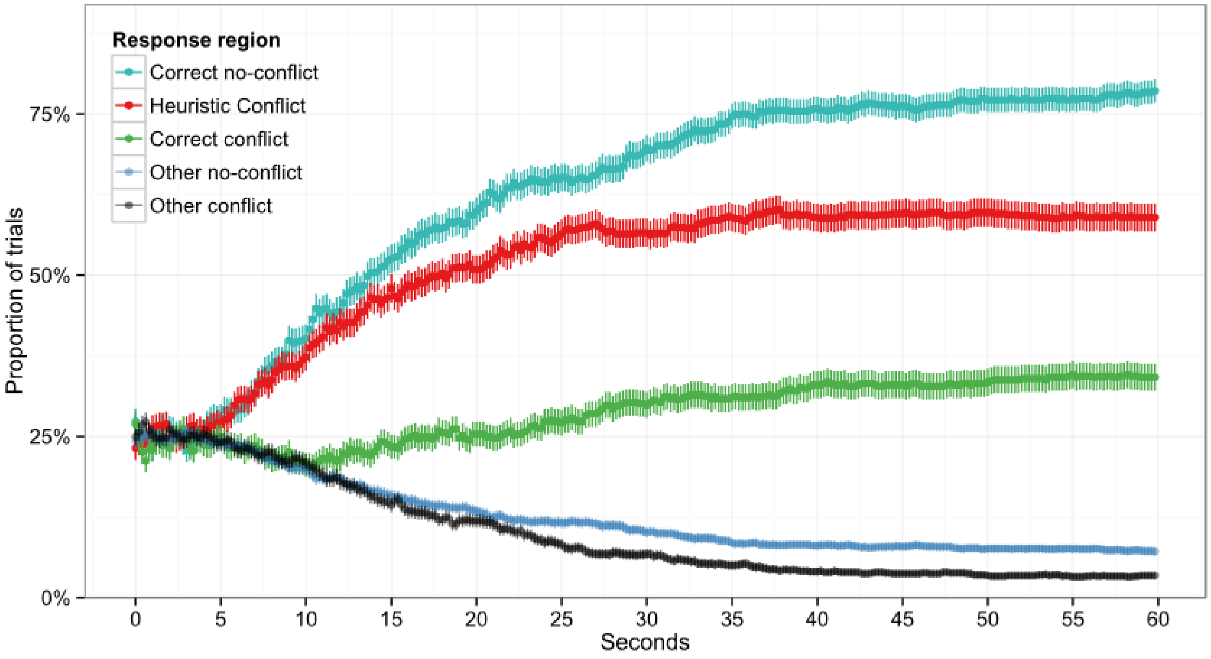


Figure 3: Proportion of mouse cursors in the region of the screen corresponding to each response options, over time, for conflict and no-conflict problems.

Figure 3 shows, for each response region, the proportion of trials in which the cursor is in that region, for both conflict and no-conflict problems. While the proportions at 60 seconds largely reflects participants' ultimate responses, earlier proportions show how these preferences developed over time. Both correct responses to no-conflict problems and heuristic responses to conflict problems were intuitively appealing, and participants began to move towards both options from after 5 seconds. After approximately 10 seconds, participants also began to move towards the correct response option on conflict problems, and the accumulation of cursors in the region of the heuristic option under conflict slowed accordingly. The proportion of cursors in the region of foil response options declined steadily in both conditions. Note that the proportions for the foil response options are averaged across the two foil options on conflict problems, and three options on no-conflict problems.

The time course data also allow us to supplement the response time analyses reported above by looking at the speed at which participants moved the mouse cursor to the region of the response option they eventually did select. Figure 4 shows this measure for correct responses to no-conflict problems, and for both heuristic and correct responses to conflict problems. The curve for each response region over time was modelled using third-order polynomial logistic regression models (or *growth curves*; see Mirman, 2014), such that the log odds of the cursor being in that region were given as $\alpha + \beta_1 time + \beta_2 time^2 + \beta_3 time^3$. Natural polynomials were used, meaning that the intercept corresponded to the log odds at 0 seconds, the linear term to the simple change over time, and the quadratic and cubic terms to higher-order differences later in the time course.[3]

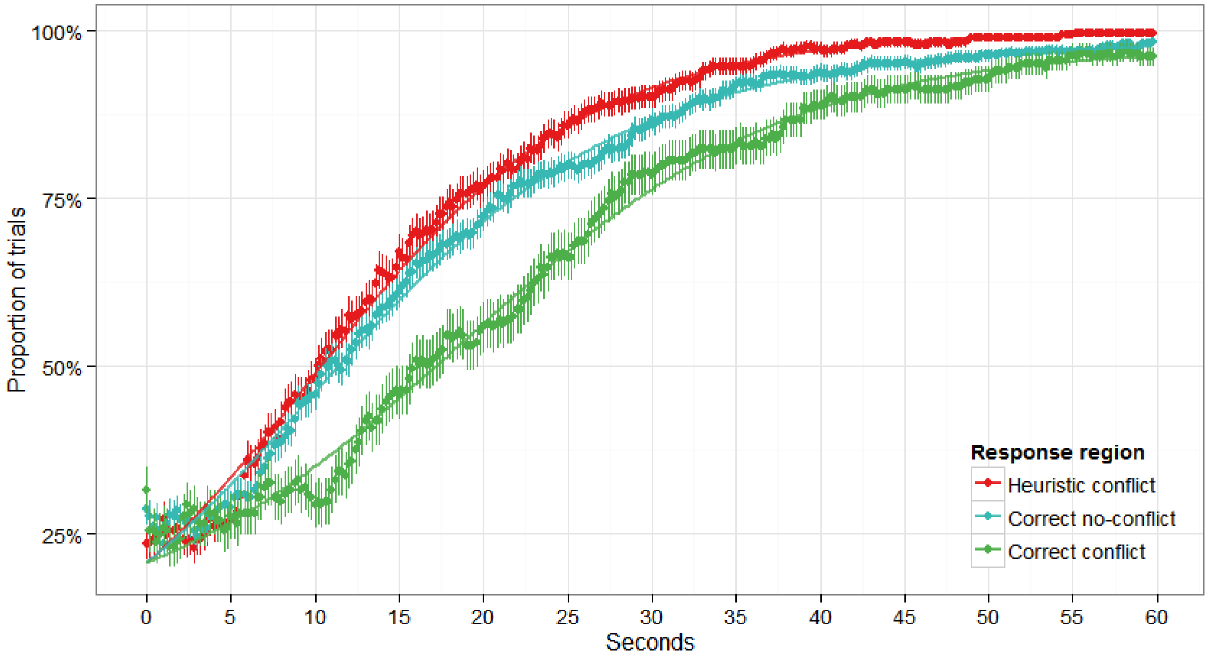


Figure 4: Proportion of mouse cursors in the region of the response option which was ultimately selected on that trial.

To test for a significant difference between two curves, a null model, in which the weights were the same for each curve, was compared with a full model, in which there were

[3] One disadvantage of using these natural polynomial terms is that they are by definition correlated, and so our model suffers from mild multicollinearity, which leads to some loss of statistical power. However, as the alternative, orthogonal polynomial terms would be difficult to interpret individually, we believe this approach lends itself to a clearer description of our data.

different β weights for each curve. Chi-squared tests were used to compare the deviance of each model, with degrees of freedom corresponding to the number of β weights added in the full model. Note that α, the intercept, was kept constant throughout. Finally, a random effect for the linear *time* term was included for each participant, to allow for individual variability in how quickly each participant moved towards a response in general. Random effects on other terms, by participant, or by problem, were considered, but led to convergence issues, and so only this term, which was found to account for the most variance, was included.

Mirroring the response time analyses, and as predicted by all dual process accounts, on conflict problems participants were faster to move towards the heuristic response option when selecting it than to the correct option when selecting it (χ^2= 4515.7, DF = 3, p < .0001), with the curves differing significantly on the linear, quadratic, and cubic terms (z’s > 5, p’s < .0001). Again consistent with the response time analyses, and contrary to previous findings supportive of the intuitive logic model, participants were faster to move towards the heuristic response on conflict problems then to move towards the correct response on no-conflict problems (χ^2, DF = 3, p < .0001), with this effect mainly driven by a significant difference on the linear term between the curves (z = 2.352, p = .0187).

In order to test for attraction towards the heuristic option on conflict trials in which the correct option was chosen, we compared the probability over time of the cursor being in the region of the heuristic option with the average probability of it being in the region of either foil option on those trials (Figure 5). A higher probability of being in the region of the heuristic option than the foils constitutes evidence of an attraction towards that heuristic response, and visual inspection of Figure 5 shows that this is the case from approximately 10 seconds onwards. Again, third order polynomial regression models were fit to this data, which showed that the difference between the curves was statistically significant (χ^2= 428.2, DF = 3, p < .0001), with significant differences on the linear, quadratic, and cubic terms (z's > 2.1, p's < .05). Therefore, when selecting the correct response, participants were slower to move away from the heuristic option than to move away from the foils, as predicted both by default-interventionist and parallel-competitive accounts.

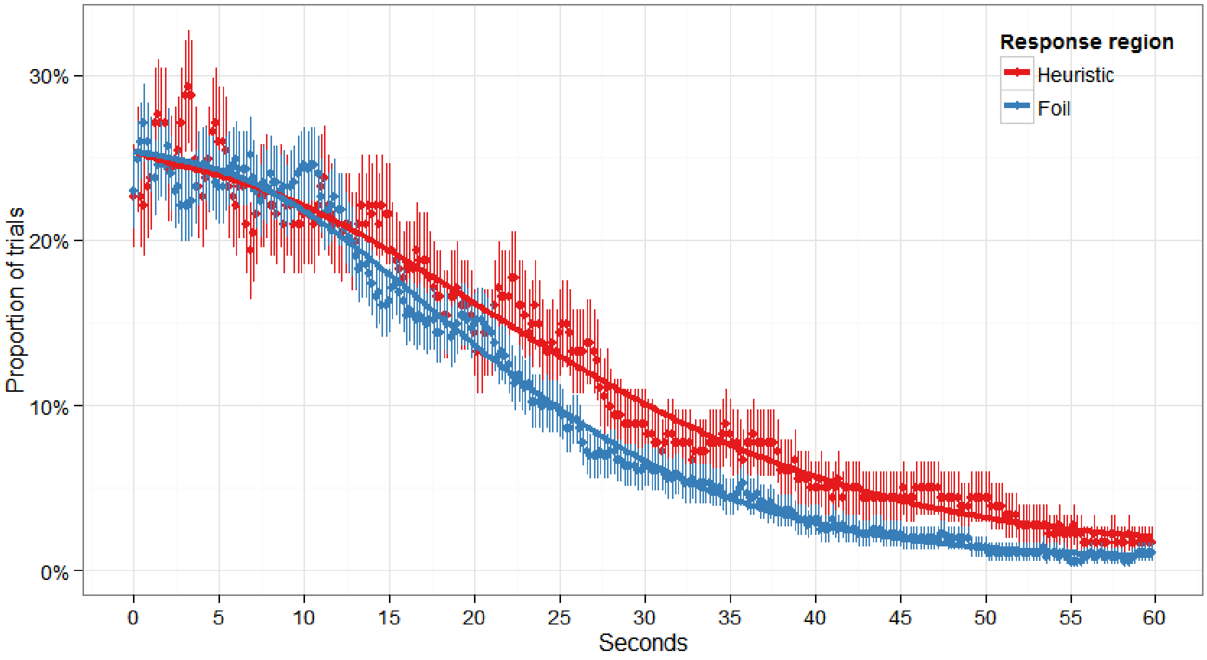


Figure 5: Proportion of trials in the region of each option, over time, for conflict problems where the correct option was eventually chosen. Error bars show standard error of measurement. Lines show fitted polynomial regression curves. Participants were more likely to be in the region of the heuristic response from around 10 seconds onwards.

A more interesting comparison is between the attraction towards the correct response option, and that towards the foil option, on conflict trials where the heuristic response is given. According to the default-interventionist account, Type 2 processes have not become engaged at this point, and so the correct response option should not be any more attractive than the foil options. According to the parallel-competitive account, on the other hand, both Type 1 and Type 2 processes should be engaged on such trials, and so participants should be drawn towards giving the response cued by Type 2 processes (that is, the correct response). Either result could be consistent with the intuitive logic theory, depending on the mechanism by which conflict is actually detected. If conflict detection occurs because Type 1 processes simultaneously cue both the correct and heuristic responses, then attraction towards the correct response option should be seen here. Conversely, if conflict is detected without Type 1 processes actually producing the correct response, then the intuitive logic theory, like the classic default-interventionist account, would predict no attraction towards the correct response option here.

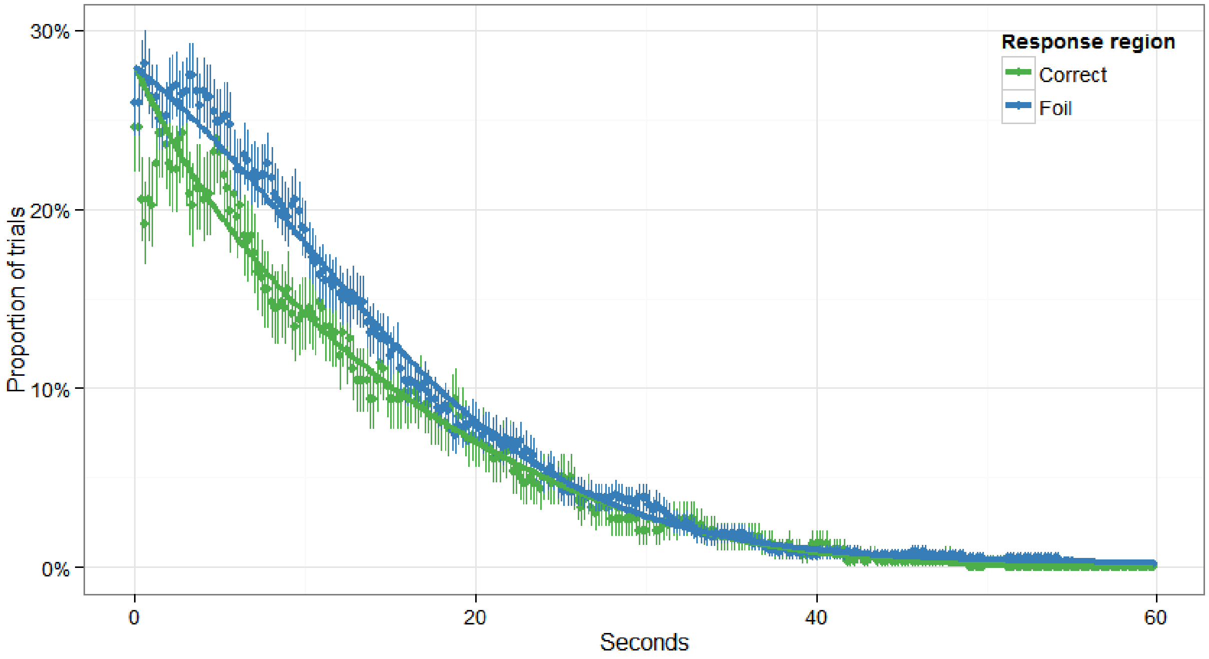


Figure 6: Proportion of trials in the region of each option, over time, for conflict problems where the heuristic option was eventually chosen. Participants were less or equally likely to be in the region of the correct *option* than a foil throughout.

Figure 6 shows that, contrary to the prediction of the parallel-competitive model, participants are not more likely to move towards the correct response option than either of the foils before selecting the heuristic option. Participants were in fact less likely to be in the region of the correct option than the foils. The polynomial regression model showed that the difference between the curves shown was again significant (χ^2 = 208.0, DF= 3, p < .0001), with significant differences between the curves on the linear, quadratic, and cubic terms (z's > 9, p's < .0001). This result indicates that the correct responses were on average actually less attractive than the foils. This is perhaps unsurprising, given that part of the difficulty of the CRT lies in the failure of intuition to support the correct response.

4. Discussion

Our results are broadly consistent with a default-interventionist dual process theory (Evans, 2006; Kahneman & Frederick, 2005). For problems with an incorrect but intuitively appealing heuristic response, this response was given more quickly, and with less evidence of conflict, than the correct response. Participants' mouse cursors began to move systematically to the region of the heuristic response option within approximately 5 seconds, compared to 10 seconds for movements to the correct response option, and this trend was evident both when analysing all trials, and trials in which the response in question was given.

When participants did give the correct response on conflict problems, they spent more time in the region of the heuristic response option than either of the foil options before doing so – a finding consistent with both default-interventionist and parallel-competitive accounts, suggesting that these participants considered the heuristic response before they reached the correct one. This finding is also consistent with modelling work (Böckenholt, 2012; Campitelli

& Gerrans, 2013) and individual differences studies (Liberali, Reyna, Furlan, Stein, & Pardo, 2011) which have shown that inhibition of the heuristic response is an important predictor of accuracy on the CRT. However, contrary to the prediction made from a parallel-competitive dual process theory (Sloman 1996; 2014), on trials where the heuristic response was given participants' were no more likely to place the cursor in the region of the correct response option than either foil option.

These results also have implications for the logical intuitions theory (De Neys, 2012; 2014). In support of this theory, a number of previous studies using simpler reasoning tasks have found that heuristic responses to conflict problems take longer than correct responses to no-conflict problems, despite both being cued by Type 1 processes (e.g. De Neys & Glumicic, 2008; Stupple & Ball, 2008). To our knowledge, the current study is the first to report response times for conflict and no-conflict versions of the CRT, and although analysis of response times was not the main focus of the current experiment, we did not find the effect that has been obtained on simpler tasks. In fact, when analysing participants' speed of movement to the response option they ultimately selected, a more sensitive measure, we found the opposite effect: participants who chose the heuristic option under conflict moved faster to their chosen option than did participants who chose the correct option in the absence of conflict. These findings held regardless of individual differences in the tendency to give the heuristic response. Thus, unlike a number of studies using simpler reasoning problems, we found no evidence that participants were slower to give intuitively-cued responses which were wrong than intuitively-cued responses which were right. Furthermore, as discussed above, we found no evidence of an attraction towards the correct response option on conflict problems where the heuristic response was given.

Previous intuitive logic studies of the CRT (De Neys et al.,2013; Gangemi et al., 2014) notwithstanding, much evidence for the theory has come from experiments with simpler tasks, such as simple syllogistic reasoning (Morsanyi & Handley, 2012), or the forced-choice base rate neglect paradigm (De Neys & Glumicic, 2008). It is possible that there are boundary conditions on the effects that have been found in these earlier experiments, and indeed this possibility has been noted by De Neys (2012; 2014). For instance, it has been demonstrated that participants report "liking" syllogisms which are logically valid more than those which are invalid, even when not asked to evaluate their logical status (Morsanyi & Handley, 2012), but also that this effect only holds for simpler logical forms (Klauer & Singmann, 2013). The operations required to reach the correct answer to our CRT problems are considerably more complex than those needed to evaluate a simple syllogism, or apply basic statistical principles. Therefore, while we do not find evidence that Type 1 processes automatically generate correct responses on the CRT, this does not rule out that they can generate correct responses on simpler tasks. Future work might use the mouse tracking paradigm to explore the role of implicit conflict detection in some of these simpler tasks.

To maximize statistical power, we analyzed performance on an extended eight-item version of the CRT. However, the two previous studies of intuitive logic in the CRT (De Neys et al., 2013; Gangemi et al., 2014), which found reduced confidence for conflict versus no-conflict problems, used only the well-known bat-and-ball problem. While our results were broadly consistent across problems, we did find some evidence that the bat-and-ball problem follows the predictions of the intuitive logic model more than the other seven problems (see the Supplementary Materials). Therefore, this discrepancy may go some way towards explaining the differences between our results and previous work, although it remains to be seen *why* the bat-

and-ball problem should behave differently to the other CRT problems. However, our main finding – that participants were not drawn towards the correct option when giving the heuristic response – was consistent across all items.

A possible criticism of our interpretation here is that a failure to confirm the predictions of a parallel-competitive model, or of one version of the intuitive logic account, should not lead us to revise our beliefs about either account. It may be the case, according to this line of reasoning, that participants are drawn towards the correct option on trials where they give the heuristic response, or that participants are more conflicted when their heuristic responses are wrong than when they are right, but we are unable to detect these mental states using our paradigm. However, it is certainly not the case that our paradigm is totally insensitive. It did reveal, for instance, that participants were initially drawn towards the heuristic option before giving the correct response on conflict problems, consistent with multiple dual process accounts. That said, it may still be the case that the attraction effects predicted by parallel-competitive and intuitive logic accounts differ in some way from the observed attraction towards the intuitively appealing response. It is not clear at this point, however, why this should be. One possibility, raised by a reviewer, is that participants who select the correct response engage in more Type 2 processing that those selecting the heuristic response, and that this could lead to a closer correspondence between their mental states and their mouse movements. Clearly, further work is needed to address these questions. Interestingly, the actual effect found was in the opposite direction to that predicted by these accounts. However, this result was unexpected, and to our knowledge is not predicted by any existing account of the task.

Of course, all of the above assumes a dual process interpretation of the CRT, as most treatments of the task do. Even in accounts which focus instead on dispositional factors (i.e.

Campitelli & Gerrans, 2014; Campitelli & Labollita, 2010), it is acknowledged that responding correctly typically requires the inhibition of the heuristic response. While we are unaware of any accounts of the CRT which do not rely on inhibition, we cannot rule out the possibility of such explanations being offered in future. The current paper, however, provides an additional constraint on such accounts, in that they should ideally predict not only observed choices, but also the process level patterns reported here. More generally, we believe the current study illustrates the value of testing theories of cognition not only against participants' final choices, but also against what we can measure of processing during the experimental task (see De Neys, 2009; Schulte-Mecklenbeck, Kuehberger, & Ranyard, 2010).

The particular application of the mouse tracking paradigm used here is a novel one, and like all experimental paradigms it rests on certain assumptions: in this case, that the movement of the mouse cursor reflects the real-time development of preference, or, in other words, that participants are more likely to move the cursor towards an option if they are considering selecting it. However, while this temporally-extended form of mouse tracking is new to psychology, researchers interested in practical human-computer interaction problems make extensive use of a similar paradigm, recording mouse movements as users interact with a graphical interface, such as a search engine results pages (see Huang, White, & Dumais, 2011). Combining this approach with eye tracking, Rodden, Fu, Aula, & Spiro (2008) report that while mouse movements correlate with gaze, they also are used in more task-specific ways, such as hovering near potential selections as a marker while eye gaze is used to explore other less likely candidates. We believe, therefore, that our data do reflect the development of participants' preferences across time.

Since its introduction in 2005, the CRT has been hugely popular as a measure of individual differences in thinking, despite only little evidence as to what underlies performance on the task. Our results go some way towards filling this gap, and suggest that responding correctly does require the activation of otherwise dormant type 2 processes to override incorrect intuitions. Future work might address the relationship between conflict on this task and individual differences. Stanovich and West (2008) proposed that normative decision making requires (1) awareness of the limitations of intuition; (2) desire to overcome those limitations; (3) inhibition of the intuitive response; and (4) ability to generate the correct response. Each of these requirements is a distinct reason for failure to produce the correct response on the CRT, and each should produce a distinctive pattern in mouse cursor movement data.

To conclude, we recorded participants' mouse cursor movements over a considerable period of time while they reasoned about CRT problems. Trends in these movements were consistent with a default-interventionist dual process theory of reasoning, where participants are initially drawn towards heuristic responses only, but in some cases engage further effortful processing to find the correct solutions. We did not find evidence that participants were drawn to correct responses on trials where these responses were not actually given, inconsistent with a parallel-competitive dual process account. Finally, contrary to previous work using simpler reasoning tasks, and confidence ratings collected on the CRT, we found no evidence that participants were conflicted when giving incorrect heuristic responses.

Acknowledgements

We wish to thank Kinga Morsanyi for providing a copy of stimuli used in Primi et al (2015), and for helpful comments. Eoin Travers is supported by a postgraduate studentship from the Department of Enterprise and Learning in Northern Ireland.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. doi:*10.1016/j.jml.2007.12.005*

Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *The Quarterly Journal of Experimental Psychology Section A*, *56*(6), 1053–1077. doi:*10.1080/02724980244000729*

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi:*10.1016/j.jml.2012.11.001*

Böckenholt, U. (2012). The Cognitive-Miser Response Model: Testing for intuitive and deliberate Reasoning. *Psychometrika*, *77*(2), 388–399. doi:10.1007/s11336-012-9251-y

Campitelli, G., & Gerrans, P. (2013). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, *42*(3), 434–447. doi:*10.3758/s13421-013-0367-9*

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making*, *5*(3), 182–191.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making*, *4*(1), 20–33.

De Neys, W. (2009). Beyond response output: More logical than we think. *Behavioral and Brain Sciences*, *32*(01), 87–88. doi:*10.1017/S0140525X09000326*

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38. doi:*10.1177/1745691611429354*

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, *20*(2), 169–187. doi:10.1080/13546783.2013.854725

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, *6*(1), e15954. doi:*10.1371/journal.pone.0015954*

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299. doi:*10.1016/j.cognition.2007.06.002*

De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: autonomic arousal and reasoning conflict. *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 208–216. doi:*10.3758/CABN.10.2.208*

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

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think When our brains detect that we are biased. *Psychological Science*, *19*(5), 483–489. doi:*10.1111/j.1467-9280.2008.02113.x*

Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review*, *13*(3), 378–395. doi:*10.3758/BF03193858*

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278. doi:*10.1146/annurev.psych.59.103006.093629*

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

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*(6), 939–946. doi:*10.1177/0956797612464058*

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

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*. doi:*10.3389/fpsyg.2011.00059*

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—in search of a phenomenon. *Thinking & Reasoning*, *21*(4), 383–396. doi:*10.1080/13546783.2014.980755*

Huang, J., White, R. W., & Dumais, S. (2011). No clicks, no problem: Using cursor movements to understand and improve search. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1225–1234). New York, NY, USA: ACM. doi:*10.1145/1978942.1979125*

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

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 267–293). New York, NY, US: Cambridge University Pres.

Klauer, K. C., & Singmann, H. (2013). Does logic feel good? Testing for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(4), 1265–1273. doi:/*10.1037/a0030530*

Koehler, D. J., & James, G. (2010). Probability matching and strategy availability. *Memory & Cognition*, *38*(6), 667–676. doi:*10.3758/MC.38.6.667*

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). lmerTest: Tests in linear mixed effects models (Version 2.0-29). Retrieved from https://cran.r-project.org/web/packages/lmerTest/index.html

Mata, A., Fiedler, K., Ferreira, M. B., & Almeida, T. (2013). Reasoning about others' reasoning. *Journal of Experimental Social Psychology*, *49*(3), 486–491. doi:*10.1016/j.jesp.2013.01.010*

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*, *27*, 1–11. doi*:10.1080/20445911.2014.986487*

Mirman, D. (2014). *Growth curve analysis and visualization using R. Boca Raton, FL: Chapman & Hall/CRC.*

Morsanyi, K., Busdraghi, C., & Primi, C. (2014). Mathematical anxiety is linked to reduced cognitive reflection: a potential road from discomfort in the mathematics classroom to susceptibility to biases. *Behavioral and Brain Functions*, *10*(1), 31. doi*:10.1186/1744-9081-10-31*

Morsanyi, K., & Handley, S. J. (2012). Logic feels so good-I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(3), 596–616. doi:*10.1037/a0026099*

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, *72*(1), 147–152. doi:*10.1016/j.jebo.2009.04.018*

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, *123*(3), 335–346. doi:*10.1016/j.cognition.2012.03.003*

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

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*, Advance online publication., *doi:10.1002/bdm.1883*

Rodden, K., Fu, X., Aula, A., & Spiro, I. (2008). Eye-mouse coordination patterns on web search results pages. In *CHI '08 extended abstracts on human factors in computing systems* (pp. 2997–3002). New York, NY, USA: ACM. doi:*10.1145/1358628.1358797*

Rolison, J. J., Evans, J. St. B.T., Walsh, C. R., & Dennis, I. (2011). The role of working memory capacity in multiple-cue probability learning. *Quarterly Journal of Experimental Psychology*, 64, 1494-1514. doi:*10.1016/j.obhdp.2012.03.003*

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110–114. doi:*10.2307/3002019*

Schulte-Mecklenbeck, M., Kuehberger, A., & Ranyard, R. (2010). A handbook of process tracing methods for decision research: A critical review and user's guide. Psychology Press.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in god. *Journal of Experimental Psychology: General*, *141*(3), 423–428. doi:*10.1037/a0025391*

Sherman, J. W., Gawronski, B., & Trope, Y. (2014). *Dual-process theories of the social mind*. Guilford Publications.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3. doi*:10.1037/0033-2909.119.1.3*

Sloman, S. A. (2014). Two systems of reasoning: An update. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind* (pp. 69–79). New York, NY, US: Guilford Press.

Spivey, M. J., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(29), 10393–10398. doi:*10.1073/pnas.0503903102*

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672–695. doi*:10.1037/0022-3514.94.4.672*

Stupple, E. J. N., & Ball, L. J. (2008). Belief–logic conflict resolution in syllogistic reasoning: Inspection-time evidence for a parallel-process model. *Thinking & Reasoning*, *14*(2), 168–181.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289. doi:*10.3758/s13421-011-0104-1*

Appendix: Cognitive Reflection Test Questions

	Conflict			No-conflict		
1	A bat and a ball together cost £1.10. A bat costs £1 more than a ball. How much does a ball cost?			A bat and a ball together cost £1.05. A bat costs £1. How much does a ball cost?		
	Correct response:	5p	(15%)	Correct response:	5p	(97%)
	Heuristic response:	10p	(83%)	Foil response:	10p	(0%)
	Foil response:	15p	(0%)	Foil response:	15p	(1%)
	Foil response:	90p	(2%)	Foil response:	90p	(1%)
2	It takes 5 machines 5 minutes to make 5 widgets. How many minutes would it take 100 machines to make 100 widgets?			It takes a machine 5 minutes to make 5 widgets. How many minutes would it take the machine to make 100 widgets?		
	Correct response:	5	(24%)	Correct response:	100	(83%)
	Heuristic response:	100	(69%)	Foil response:	5	(2%)
	Foil response:	50	(4%)	Foil response:	50	(13%)
	Foil response:	10	(3%)	Foil response:	10	(2%)
3	In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how many days would it take for the patch to cover half of the lake?			In a lake, there is a patch of lily pads. Every day, the patch grows by 10m^2. If it takes 48 days for the patch to cover the 150m^2, how many days would it take for the patch to cover 140m^2?		
	Correct response:	47	(26%)	Correct response:	47	(79%)
	Heuristic response:	24	(58%)	Foil response:	24	(17%)
	Foil response:	12	(15%)	Foil response:	12	(4%)
	Foil response:	2	(2%)	Foil response:	2	(0%)
4	If you flipped a fair coin twice, what is the probability that it would land 'Heads' at least once?			If you flipped a fair coin twice, what is the probability that it would land 'Heads' exactly once?		
	Correct response:	75%	(4%)	Correct response:	25%	(68%)
	Heuristic response:	50%	(84%)	Foil response:	50%	(26%)
	Foil response:	25%	(11%)	Foil response:	75%	(6%)
	Foil response:	100%	(1%)	Foil response:	100%	(0%)
5	If 3 elves can wrap 3 toys in 1 hour, how many elves are needed to wrap 6 toys in 2 hours?			If 3 elves can wrap 3 toys in 1 hour, how many toys could 6 elves wrap in half an hour?		
	Correct response:	3	(73%)	Correct response:	3	(71%)
	Heuristic response:	6	(21%)	Foil response:	6	(20%)
	Foil response:	1	(2%)	Foil response:	1	(1%)
	Foil response:	12	(4%)	Foil response:	12	(8%)
6	Ellen and Kim are running around a track. They run equally fast but Ellen started later. When Ellen has run 5 laps, Kim has run 10 laps. When Ellen has run 10 laps, how many has Kim run?			Ellen and Kim are running around a track. They started at the same time, but Kim is twice as fast as Ellen. When Ellen has run 5 laps, Kim has run 10 laps. When Ellen has run 10 laps, how many has Kim run?		
	Correct response:	15	(73%)	Correct response:	20	(98%)
	Heuristic response:	20	(27%)	Foil response:	15	(2%)
	Foil response:	5	(0%)	Foil response:	5	(0%)
	Foil response:	19	(0%)	Foil response:	19	(0%)

7	Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are there in the class?			Jerry received both the 2nd highest and the 2nd lowest mark in the class. How many students are there in the class?		
	Correct response:	29	(26%)	Correct response:	3	(79%)
	Heuristic response:	30	(72%)	Foil response:	2	(13%)
	Foil response:	40	(2%)	Foil response:	5	(8%)
	Foil response:	5	(0%)	Foil response:	10	(0%)
8	In an athletics team tall members tend to win three times as many medals than short members. This year the team has won 60 medals so far. How many of these have been won by short athletes?			In an athletics team tall members tend to win twice as many medals than short members. This year the team has won 60 medals so far. How many of these have been won by short athletes?		
	Correct response:	15	(44%)	Correct response:	20	(58%)
	Heuristic response:	20	(52%)	Foil response:	15	(12%)
	Foil response:	30	(1%)	Foil response:	30	(26%)
	Foil response:	50	(3%)	Foil response:	50	(4%)

Note. Percentages in parentheses show the proportion of participants who gave each response.