Memory & Cognition (2022) 50:348–365
https://doi.org/10.3758/s13421-021-01224-8

Eye tracking and the cognitive reflection test: Evidence for intuitive correct responding and uncertain heuristic responding

Zoe A. Purcell[1] · Stephanie Howarth[2] · Colin A. Wastell[1] · Andrew J. Roberts[2] · Naomi Sweller[1]

Accepted: 27 July 2021 / Published online: 13 August 2021

Abstract
The Cognitive Reflection Test (CRT) has been used in thousands of studies across several fields of behavioural research. The CRT has fascinated scholars because it commonly elicits incorrect answers despite most respondents possessing the necessary knowledge to reach the correct answer. Traditional interpretations of CRT performance asserted that correct responding was the result of corrective reasoning involving the inhibition and correction of the incorrect response and incorrect responding was an indication of miserly thinking without feelings of uncertainty. Recently, however, these assertions have been challenged. We extend this work by employing novel eye-tracking techniques to examine whether people use corrective cognitive pathways to reach correct solutions, and whether heuristic respondents demonstrate gaze-based signs of uncertainty. Eye movements suggest that correct responding on the CRT is the result of intuitive not corrective cognitive pathways, and that heuristic respondents show signs of gaze-based uncertainty.

Keywords Cognitive Reflection Test · Dual process · Eye tracking · Conflict · Uncertainty

Introduction

Dual process theories of reasoning distinguish between Type 1 (implicit and automatic) and Type 2 (deliberative and reflective) thinking. There are various dual process models that contain distinct proposals about the sequence of cognitive events that occur from when a reasoner encounters a problem to when they reach a solution, and the role of uncertainty within that sequence. These distinctions are most evident when comparing the default-intervention (e.g., Evans & Stanovich, 2013; Kahneman, 2011) and hybrid dual process models (e.g., De Neys, 2012, 2014; Pennycook et al., 2015; Thompson et al., 2011).

The following sections focus on the different proposals put forward in the default-intervention and hybrid dual process models, first, as they relate to cognitive pathways and second, as they relate to the role of uncertainty. We use the term 'cognitive pathway' to refer to the sequence of cognitive events occurring during reasoning and, in particular, the order in which information is processed. The term 'uncertainty' refers to the phenomenon in which reasoners register at some level that their solution is not completely warranted – a factor that hybrid dual process models of reasoning have suggested plays an important role in the engagement of deliberative thinking (for a review, see De Neys, 2018).

To narrow the scope of the current article, we focus on cognitive pathways and uncertainty in relation to the CRT. The CRT is one of many 'bias tasks' that are at the centre of empirical investigations of human reasoning and dual process theories. The first CRT item states: "A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" The correct answer is 5 cents, however, the heuristic answer that comes to mind for most respondents is 10 cents (e.g., Bago et al., 2019; Frederick, 2005). The CRT has reached a point of empirical maturity that facilitates a substantial exploration of the theoretical and empirical considerations relating to cognitive pathways and uncertainty. These considerations are explored in more detail in the following sections.

✉ Zoe A. Purcell
zoe.purcell@iast.fr

[1] Department of Psychology, Macquarie University, Sydney, NSW, Australia

[2] Department of Cognitive Science, Macquarie University, Sydney, NSW, Australia

Cognitive pathways and the CRT

Theories of corrective and intuitive pathways

A key difference between the default-intervention model and the hybrid models are the proposed cognitive pathways that lead to heuristic versus logical responses in cases where the two solutions differ (see Bago & De Neys, 2017). The default-intervention model proposed that heuristic responses are cued in Type 1 thinking but can then be overridden and corrected by Type 2 processes. The generation of an incorrect heuristic response, and subsequent overriding of that response with the correct solution, can be interpreted as a *corrective* cognitive pathway.

In contrast, the hybrid models assert that multiple Type 1 processes, including heuristic and logical responses, can be triggered simultaneously, due to most adult reasoners having elementary intuitions about logical and probabilistic principles. Although additional Type 2 processing may be required to complete more complex processes (see De Neys, 2018), the hybrid models emphasise the possibility for people to reach logical solutions via Type 1 processes (e.g., De Neys, 2012, 2014; Pennycook et al., 2015; Thompson et al., 2011). Reaching a logical solution via Type 1 processing can be interpreted as an *intuitive* cognitive pathway; that is, when a person automatically begins their problem solving with processes leading to the correct solution.

To illustrate, consider the bat and ball problem introduced earlier. A person using a corrective pathway would first consider the heuristic '10 c' response and then correct that response to reach the solution '5 c'. Alternatively, a person using an intuitive pathway would consider the final, correct solution '5 c' from the outset, without first considering the heuristic response of '10 c' and overriding it using deliberative thinking.

Evidence of corrective and intuitive pathways

The CRT was originally developed as a measure of reflection; that is, one's ability to suppress an intuitively appealing but incorrect answer (i.e., the Type 1, heuristic response) and substitute it with the correct response (i.e., the Type 2, logical response; Frederick, 2005). This interpretation assumed that corrective pathways were used to reach logical solutions on the CRT. However, since the original CRT publication, several studies have directly examined the corrective interpretation using techniques such as mouse tracking (Travers et al., 2016), think-aloud protocols (Szaszi et al., 2017), and mathematical experience manipulations (Purcell et al., 2021).

Travers et al. (2016) examined the mouse trajectories of participants solving the CRT. In line with the default-intervention model, they found that correct respondents were tempted by the heuristic response before selecting the correct option, whereas heuristic respondents were not tempted by the correct response prior to selecting the heuristic option. It can be inferred from this finding that corrective cognitive pathways may have been employed because a participant using corrective reasoning would first be tempted by the heuristic option before selecting the correct option.

However, the aggregation of data across participants prevented researchers establishing whether this reflects outliers or a typical pattern of responding, or whether some participants may have been able to reach the correct response without considering the heuristic option first. Moreover, although mouse-tracking has been used to examine decision strategies for other reasoning tasks (Szaszi et al., 2018), there is evidence that such paradigms may not capture automatic or implicit processes (Glöckner & Betsch, 2008; Glöckner & Herbold, 2011). Mouse-trajectories are often correlated with eye-movements, and as such offer an indication of the pieces of information that an individual is considering. However, the direct observation of eye movements would provide a more rigorous examination of the information considered by the reasoner (Ball et al., 2003; Thompson, 2021).

Szaszi et al. (2017) examined reasoning on the CRT using a think-aloud protocol. The participants verbally conveyed their reasoning to the experimenter as they completed the task. Of the correct respondents, only 23% began thinking aloud about the heuristic option. The majority (77%) began thinking aloud about the correct option. In another experiment, Mata et al. (2013; Study 5) found that 28% of correct respondents indicated, after responding to the task, that the heuristic response had "come to mind". We can infer from these findings that most correct respondents in the Szaszi et al. (2017) and Mata et al. (2013, Study 5) studies were using *intuitive* not *corrective* cognitive pathways.

However, think-aloud protocols only allow for the analysis of explicit cognitive pathways, that is, those we have conscious access to and can recall and communicate (Crutcher, 1994). Despite the individual-level analysis indicating the use of mostly *intuitive* pathways, it is possible that this methodology underestimates the initial activation of heuristic thought processes that would indicate the use of *corrective* pathways. Like mouse-tracking, think aloud and recall protocols offer limited information about the potential implicit components of the cognitive pathways employed on the CRT.

Overall, studies administering the CRT have found evidence primarily for the use of correct responding via intuitive reasoning (e.g., Bago & De Neys, 2019; Mata et al., 2013; Szaszi et al., 2017). However, these have been conducted at a single point in time – without manipulations of mathematical experience.

Effects of experience on cognitive pathways

The few people providing correct responses in previous studies may have had high numeracy, such that they had learned

the solution processes to automation and could respond via intuitive pathways. It has been suggested that people with greater mathematical experience, numeracy, or cognitive abilities are more likely to use autonomous, non-working-memory dependent, logical intuitions (De Neys & Pennycook, 2019; Peters, 2012; Purcell et al., 2021; Raoelison et al., 2020; Stanovich, 2018). In contrast, those with the relevant mindware to reach the correct solution but insufficient experience for automation may be more likely to use working memory-dependent, corrective pathways (see Purcell et al., 2021; Stanovich, 2018).

Mathematical experience may impact the cognitive pathway a person uses to reach correct solutions on the CRT. Therefore, the current study includes a training manipulation to gradually increase the experience and accuracy of the participants. Although the individual trajectories of working-memory dependence and hence cognitive pathways may depend on factors like numeracy[1] and working-memory capacity (Purcell et al., 2021), by increasing the potential for intra-individual variation in accuracy, we increase the likelihood of observing correct responding and corrective cognitive pathways that, to date, have remained empirically allusive.

Uncertainty and the CRT

The role of uncertainty in reasoning

Although they are often examined together, one can distinguish between two primary avenues of research regarding reasoning and uncertainty: One examining whether reasoners are miserly, biased, and unaware of their errors, and the other looking at whether and how uncertainty may be involved in the engagement of effortful thinking. Many decades of thinking research have established that human reasoning is often biased. Bias tasks like the CRT have been used to study reasoning bias because they elicit erroneous responses despite respondents often possessing sufficient mindware to reach the correct solution. However, recent evidence has emerged suggesting that erroneous respondents may not be as unaware of the tasks' logical principles as previously thought (e.g., De Neys & Glumicic, 2008; Pennycook et al., 2015; Thompson & Johnson, 2014).

Authors have also suggested that cognitive uncertainty may play a role in triggering effortful, deliberate thinking; however, they emphasise different specifications for the mechanisms that underlie that uncertainty (De Neys, 2012, 2014; Stanovich, 2018;Thompson et al., 2011 ; Thompson & Morsanyi, 2012). Thompson and colleagues suggested that the fluency of the response, influenced by factors such as the ease with which a response comes to mind, impacts our "feeling of rightness" (Thompson et al., 2011; Thompson & Morsanyi, 2012). When the response is less fluent, the feeling of rightness is lowered, and deliberative thought is cued. The logical intuition model on the other hand asserts that multiple Type 1 processes can be initiated simultaneously, and that their relative activation can elicit uncertainty[2] (De Neys, 2012, 2014). When the initiated processes have similar levels of activation, cognitive 'conflict' between the processes is generated, and deliberative thought is engaged. The current study explores these proposals by using novel eye-tracking techniques to examine whether heuristic CRT respondents show signs of uncertainty, and to explore the specificity of the mechanisms that underlie that uncertainty.

The evidence for uncertainty and its specification

Previous studies have examined whether respondents providing incorrect responses on bias tasks are truly unaware of their error, or if they show some sensitivity to the conflicting logical principles. These studies have typically compared indicators of uncertainty for 'lure' and 'no lure' (control) versions of bias tasks, such as base-rate or syllogistic reasoning problems (Bago & De Neys, 2017; De Neys, 2012). Lure items are those that have different logical and heuristic responses, whereas no-lure items have the same logical and heuristic response. For example, the bat and ball problem introduced earlier has a logical answer (5 c) and a heuristic answer (10 c). In contrast, a no-lure version of this problem might state: A bat and a ball cost $1.10 together. The bat costs $1.00. How much does the ball cost? In this no-lure example, the heuristic and logical principles cue the same response (10 c).

To determine whether respondents to lure items are sensitive to the conflict between logical and heuristic response processes, uncertainty has been compared between heuristic responding on lure items (heuristic lure trials) and correct responding on no lure items (correct-no lure trials; see De Neys, 2012). Studies employing the CRT have observed differences in uncertainty between heuristic lure and correct no lure trials with many measures of uncertainty. For example, greater uncertainty on heuristic lure trials relative to correct-no lure trials has been observed consistently for indicators such as feelings of confidence (Bago et al., 2019; De Neys et al., 2013) and error (Gangemi et al., 2015), and, in some cases,

[1] For example, we might expect participants with lower mathematical experience to move from incorrect responding (e.g., answering 10 c to the bat and ball problem) to correct responding (e.g., 5 c) with effortful corrective pathways, and participants with intermediate mathematical experience to move from providing correct responses using effortful corrective pathways to providing correct response via faster intuitive pathways.

[2] We note that this conflict would not require the two (or more) competing processes to be fully generated, but that the solution processes are activated.

response latencies (Bago et al., 2019; Hoover & Healy, 2019; cf. Stupple et al., 2017). However, Travers et al. (2016) did not find evidence for differences in response latencies or cursor-based indices.

Despite some null results (Stupple et al., 2017; Travers et al., 2016), there is considerable evidence to suggest that incorrect respondents show some sensitivity to or interference from the problems' logical principles that conflict with the heuristic response (Bago et al., 2019; De Neys et al., 2013; Gangemi et al., 2015; Hoover & Healy, 2019). This has prompted investigations into the specificity of the uncertainty signals; that is, whether the uncertainty is process- or non-specific (Bago et al., 2019; Travers et al., 2016). As in Bago et al. (2019), we will use the term *process-specific* uncertainty to refer to uncertainty that reflects the competition between two (or more) specific competing solution processes, and *non-specific* uncertainty to refer to that which does not reflect the weighing of or competition between two or more specific processes.

Bago et al. (2019) used a second-guess paradigm; they first asked respondents to generate an answer to the bat and ball problem, then they asked the participants to make a second guess between multiple choice options that do not include the heuristic 10 c option. For example, participants might select between 1 c, 5 c, and 15 c. If the participant experienced a "specific conflict signal", such that they were weighing the two plausible responses (5 c and 10 c), they were expected to select the 5 c option once the 10 c option was removed. If, in contrast, the participant was unsure about their response but did not have a specific alternative response in mind, their second-guess option should be selected at random. Interestingly, participants did not opt for the correct response at their second attempt, but they did intuit that the correct answer was smaller than their original (10 c) response. Bago et al. (2019) interpreted this as an indication of medium-specific conflict signals, suggesting that reasoners are more sensitive to their errors than traditionally assumed.

Specificity can also be inferred from Travers et al.'s (2016) examination of participants' mouse trajectories on a multiple-choice version of the extended eight-item CRT (Primi et al., 2016). They found that correct respondents showed some mouse trajectory curvature towards the heuristic response (e.g., 10 c) however they did not observe heuristic respondents showing an attraction to the correct response (e.g., 5 c). This may indicate that the incorrect reasoners in their sample did not experience uncertainty, or that their uncertainty was non-specific. That is, the participants may have sensed that their 10 c response was not fully warranted but not suspected that the 5 c response could be correct. Alternatively, they may have been attracted to the 5 c option, but the mouse-tracking methodology was not sensitive enough to detect it. We continue to explore the specificity of uncertainty on the CRT by employing more sensitive eye-tracking techniques.

The current paradigm

The current paradigm builds on previous studies examining reasoning and uncertainty on the CRT by incorporating a mathematical training manipulation and novel eye-tracking techniques. We present people with CRT-like problems and four multiple-choice options and examine how much visual consideration people give to each of the multiple-choice options. We examine cognitive pathways for correct responding by comparing eye movements for correct and heuristic trials on CRT-like lure items. We examine uncertainty for heuristic responding by comparing confidence ratings and eye-movements between heuristic lure and correct-no lure trials.

The CRT is widely known for eliciting incorrect responses from most respondents. Therefore, to examine reasoning used to reach correct solutions it is important to consider paradigms that increase the likelihood of observing correct responding. This can be achieved, for example, by including many participants in the hope that a reasonable portion will exhibit correct responding, or a training manipulation to improve performance. The inclusion of a training paradigm, however, has the added benefits of increasing intra-individual variance and increasing the likelihood of observing corrective pathways. As in Purcell et al. (2021), we expected that correct responding on lure items would increase with training.[3] While the examination of reasoning prior to any training is still included in this study (refer to Test Block 1 in the following sections), the training manipulation allowed for the more rigorous examination of heuristic versus correct responding as a factor that varies for each individual throughout the study, rather than between individuals as is typical in non-training paradigms. This ability to track performance within individuals is particularly important when examining variables that can vary substantially between individuals like those based on eye tracking.

Eye tracking is a non-invasive method that can be used to examine cognitive pathways and uncertainty at the time of processing, without relying on subjective, explicit, post hoc judgments, or influencing the reasoning process. Eye movements provide information about cognitive processes that the researcher cannot determine from performance alone and that the participant may not be consciously aware of (e.g., Bruckmaier et al., 2019; Green et al., 2007; Stephen et al., 2009). There are several eye-tracking measures that can be examined to assess cognitive processes throughout reasoning. We focus on the number of *fixations*, which represent the maintenance of gaze on a certain location, and *dwell*, which represents the duration of fixations in a particular area.

[3] We do not make specific hypotheses for interaction effects between training (indicated by effects of 'test block' below) and cognitive pathways/uncertainty because this was beyond the scope of the current research questions and would require the added consideration of individual factors such as numeracy and intelligence.

Fixations can be examined to determine which pieces of information a person considers (e.g., Ball et al., 2006) and dwell can be examined to assess the depth with which that information is processed (e.g., Glöckner & Herbold, 2011; Rayner, 1998; Velichkovsky, 2014; Velichkovsky et al., 2002).

Research questions and predictions

The current study addressed two primary research questions. First, whether cognitive pathways differ for correct and heuristic trials on CRT-like lure items. If our results were in line with the default-intervention dual process model and Travers et al. (2016), we would expect participants giving the correct response on CRT-like items to demonstrate corrective cognitive pathways. That is, they would show evidence for having considered the heuristic response option to a greater extent than the other foil options before finally selecting the correct option. However, if our findings were in line with the hybrid dual-process models and studies demonstrating intuitive correct responding (e.g., Szaszi et al., 2017), we would expect that participants giving the correct response would not consider the heuristic option to a greater extent than the other foil options.

Our second research question investigated whether there is evidence for greater uncertainty on heuristic lure compared to correct-no lure trials, and if so, whether this uncertainty presents as a process-specific or non-specific signal? In line with previous studies (Bago et al., 2019; Hoover & Healy, 2019), we expected that uncertainty (operationalised in this experiment via confidence, fixations, and dwell) would be greater for heuristic lure trials than correct-no lure trials. We did not have strong hypotheses regarding the specificity of that uncertainty as a process- or non-specific signal, or whether that specificity would change as training increased.

To address our second research question, we first establish that greater information search, a sign of uncertainty, had taken place during heuristic lure trials than correct-no lure trials, and then examine eye movements to determine the information that was processed and the specificity of the uncertainty. To illustrate, imagine a respondent gives the heuristic (10 c) response to the bat and ball problem. If they display low uncertainty in their final response selection (10 c) they would be expected to give little consideration to their non-selected options (5 c or other foil options). In contrast, a respondent with high uncertainty would be expected to show greater consideration of the alternative, non-selected responses.

We then extend this examination of eye movements to look for evidence of process-specific or non-specific uncertainty. To illustrate with the same example, a person with high process-specific uncertainty would be expected show greater consideration of the non-selected logical (5 c) response than the other non-selected responses. Alternatively, a respondent with high non-specific uncertainty would not necessarily favour the non-selected logical solution (5 c) over the other non-selected responses.

Method

Participants and design

A 4 (test block: T1, T2, T3, T4) × 2 (problem type: lure and no lure) within-subject design was used. Participants were randomly allocated to one of two constraint conditions; however, there were no differences in performance, cognitive pathways, or uncertainty between these groups and therefore they were collapsed.[4] Participants were 38 undergraduate psychology students at Macquarie University (Sydney, NSW, Australia) who were awarded course credit for participation. Participants were 27 females and 11 males with ages ranging from 18 to 36 (M = 19.76, SD = 3.54). The sample size was sufficient to detect similar effects to those observed in Travers et al. (2016; e^{β} = 0.92, $t(72.1) = 4.119$, $p < .0001$) for our highest order interaction of interest – a two-way interaction between AOI and trial type used to test corrective pathways (see *Results*).[5] All participants had normal vision. Figure 1 presents the timeline for the overall experiment.

Materials and apparatus

Lure and no-lure items

Forty-two items (21 lure and 21 no lure) were used. The objects and quantities were different for all items. The lure items were developed to have a heuristic response, in the format of the CRT (Frederick, 2005). The no-lure items were developed to mirror the same structure but without a heuristic incorrect response. An example of a lure item is: "It takes three spiders 3 minutes to make three webs. How long would it take 100 spiders to make 100 webs?" (correct answer: 3 min). This item has an incorrect heuristic response of "100". An example of a corresponding no-lure item is: "It takes one factory 10 days to build 20 cars. How many days would it take one factory to

[4] While completing the CRT, participants were required to memorise a grid pattern: low – 3 × 3 grid with a one-piece pattern of three coloured squares or high – 3 × 3 grid with a two-piece pattern of four coloured squares (for a similar paradigm, see Purcell et al., 2021). A programming error yielded longer display times for the grid patterns than in previous uses of this technique which may have rendered this manipulation ineffective.

[5] In the current study, this interaction was not significant. A post hoc power analysis via simulation in R indicated it was powered at 58% [*CI* 47.71, 67.80]. However, the direction of the effect supported neither the alternative hypothesis nor another sensible theoretical interpretation suggesting this null result is unlikely to stem from power issues (see *Results*).

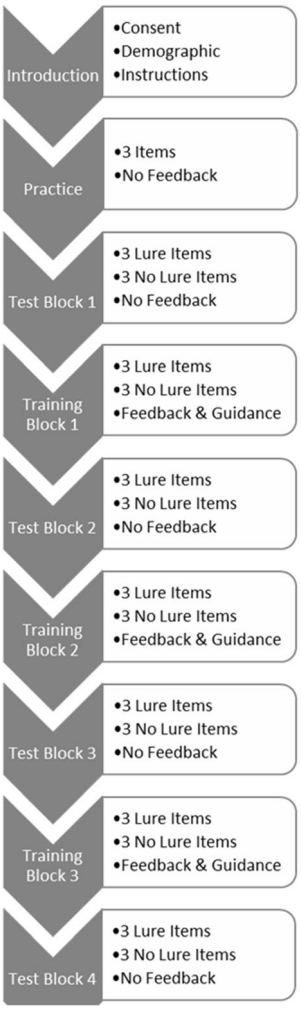

Fig. 1 Summary illustration of the experimental timeline. The order of items was randomised within blocks and counterbalanced between blocks

build 40 cars?" (correct answer: 20 days). This item has no competing heuristic response. All participants saw the same set of items and no item was repeated. A full description of the items is available here https://osf.io/ej3n2/.

The reasoning task

The reasoning task was based on a previous paradigm used to improve performance and increase variance in working memory dependence on the CRT (Purcell et al., 2021). Participants were presented with three practice items to familiarise them with the task and eye-tracking equipment. Practice items were simple no-lure problems followed by a question about their confidence in their response (see Fig. 2); they did not include feedback. Participants were then presented with the 42 lure and no-lure items described above. These were presented in seven blocks of six items in the order: test block 1 (T1), training block 1, test block 2 (T2), training block 2, test block 3 (T3), training block 3, and test block 4 (T4; see Fig. 1). Each block included one lure item and one no-lure item for each of the three CRT questions.

All items were presented in the format shown in Fig. 2. Participants were presented with the item and four multiple choice options (see Fig. 2A). Immediately after making their selected response, participants were asked about their confidence in that response (see Fig. 2B). Additionally, items in the training blocks were followed by feedback (correct or incorrect) and guidance; those who gave correct responses were provided with a short explanation of why this was the correct response, and those who gave incorrect responses received more interactive guidance, for example, learning to translate questions into algebraic forms by following several steps (for a full description of items and feedback see https://osf.io/ej3n2/). To prevent order effects, items were counterbalanced between blocks and randomised within blocks (see Appendix 1). The reasoning task took approximately 40 min.

Eye-tracking apparatus and areas of interest (AOIs)

The reasoning task was presented on a 24.5-in. LCD monitor (BenQ XL2540, refresh rate 240 Hz, natural resolution 1,920 × 1,080). The program was run on Experiment Builder presentation software 1.10.165 (SR-Research) and was entirely mouse driven to reduce eye movements off the screen (i.e., onto the keyboard). Eye movements were recorded monocularly (right eye) with a desk mounted eye-tracker sampling at a rate of 1,000 Hz (EyeLink 1000; SR Research Ltd., Osgoode, Ontario, Canada). A chinrest was used to stabilise head movements and maintain viewing distance (800 mm). Data were extracted using Eyelink Data Viewer (SR-Research) and analysed with SPSS Version 26.0.

For coding purposes, areas of interest (AOIs) were assigned to each of the four multiple-choice alternatives. These were coded relative to the response that the participant had chosen on each trial. For lure items these were 'Selected', 'Other-relevant', 'Other-1' and 'Other-2'. Therefore, for participants who gave the heuristic response, the AOI assigned to the heuristic answer was labelled 'Selected' and the AOI assigned to the correct response was labelled 'Other-relevant' and vice versa for the participants who gave the correct answer. For no-lure items, the AOIs were labelled 'Selected' and 'Other-1', 'Other-2', 'Other-3' because there was no theoretical distinction between the incorrect 'Other' AOIs. The analyses treated each AOI separately to gain a comprehensive understanding of the dispersion of fixations and dwell between the four multiple-choice options as well as the differences in fixations and dwell between each option. Each multiple-choice option was randomly allocated to a corner of the screen (see Fig. 2A). Therefore, AOIs were dynamic, such that they reflected the option value and not the placement. The

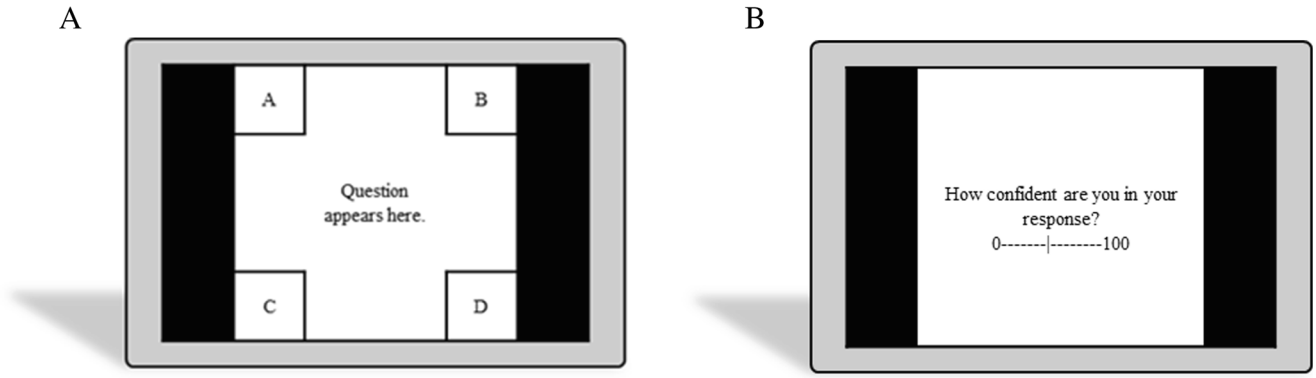

Fig. 2 (**A**) Layout used for all items. Participants were required to choose an answer from the four options in the corners. The position of the answers was randomised for each item. (**B**) Layout used for participants to rate their confidence in their response. Note. The scale has been adjusted for legibility

number of fixations and their duration (dwell) was summed for each AOI by trial.

Measuring cognitive pathways

Eye movements were compared for lure trials on which the participant gave the correct responses and those for which participants gave heuristic responses. We recorded participants' fixations and dwell that occurred within the AOIs assigned to the four multiple-choice options: Selected, Other-Relevant, Other-1, Other-2. For correct trials, the selected option was the logical response and the other-relevant option was the heuristic response. For the heuristic trials, the selected option was the tempting but incorrect response and the other-relevant option was the logical response. For all trials, we expected to observe the greatest proportion of fixation and dwell on the final, selected response. However, the participants' consideration of the alternative responses was used to assess whether their cognitive pathway leading to that final selection was corrective or intuitive. As in Travers et al. (2016), we examined whether correct responders considered the heuristic response more than heuristic responders considered the correct response (see Fig. 3). If our data reflected a corrective cognitive pathway, we would expect correct-trials to yield a greater proportion of fixation and dwell on the other-relevant option than the other-1 or other-2 (see Fig. 3(A)). If, however, our data reflected an intuitive cognitive pathway, we would expect that the selected option would have greater fixation and dwell than the other-relevant, other-1 and other-2 options (see Fig. 3(B)).

We note that the difference in the proportion of fixations and dwell between correct trials and heuristic trials predicted here are not sufficient to infer the order in which the information was processed. However, it is not necessary to examine the specific order of consideration to test whether the eye movements are in line with the default intervention model. The default intervention model suggests that the heuristic option would be considered first, and then the logical option. A crucial inference from this proposal is that the heuristic option should be considered to a greater extent than the other non-heuristic foil options. We found no evidence to suggest that this was the case and, therefore, did not pursue the examination of order effects.

Measuring uncertainty

We examined confidence- and gaze-based measures of uncertainty for participants completing the CRT. In line with previous studies, we compared uncertainty on lure problems for which heuristic responses had been made (heuristic lure trials) to uncertainty on no lure problems for which correct responses had been made (correct-no lure trials; e.g., De Neys et al., 2013). Participants reported their confidence on a scale from 0 to 100 immediately after making their selected response. These scores were reverse coded to reflect uncertainty (for similar techniques, see De Neys et al., 2011; De Neys et al., 2013; Gangemi et al., 2015).

To investigate gaze-based indicators of uncertainty, we examined the proportion of fixations and dwell across the four multiple choice options. A lower proportion of fixations and dwell on the selected response – and hence, higher inspection of the non-chosen options – was thought to indicate a greater consideration of the alternative responses during reasoning and hence greater uncertainty in the response that was finally selected. Conversely, if the proportion of fixations and dwell on the chosen response was high – and hence, lower inspection on the non-chosen options – this was thought to indicate that the other options were given relatively less consideration and hence lower uncertainty in the selected response.

Patterns of fixation and dwell on the heuristic lure items were also examined to explore whether the data indicated a process-specific or non-specific signals of uncertainty. If the

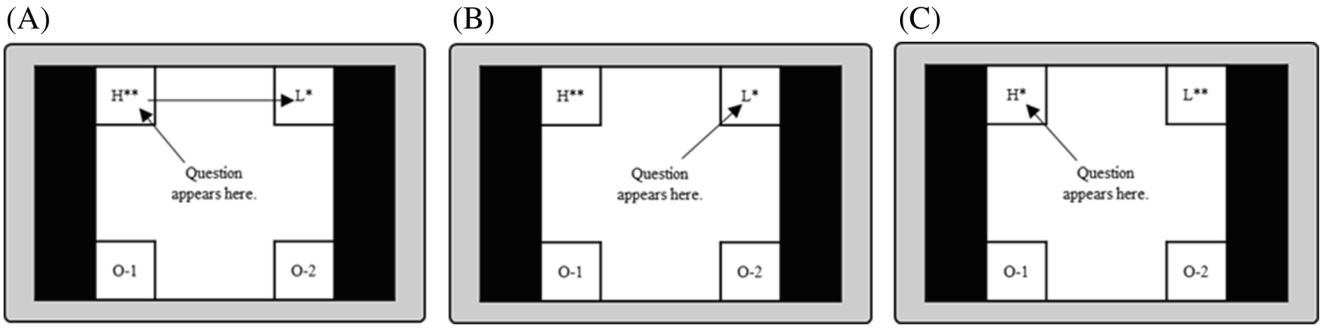

Fig. 3. Hypothetical gaze patterns for (**A**) correct trials reflecting a corrective pathway, (**B**) correct trials reflecting an intuitive pathway, and (**C**) incorrect heuristic trials. Arrows represent gaze movements starting at the question and moving to the final response selection. H = Heuristic response option, L = Logical response option, O-1 and O-2 = other incorrect response options. * Indicates the selected option, ** indicates the other-relevant option

data reflected uncertainty as a non-specific signal, we would expect that the proportion of fixations and dwell would be greatest for the selected (heuristic) response than the remaining responses and similar for all non-selected alternatives (see Fig. 4a). However, if the data reflected uncertainty as a process-specific signal, we would expect less similarity between the non-selected responses – the proportion of fixation and dwell would be greater for the non-selected correct response than the other non-selected responses (see Fig. 4b).

Procedure

Participants gave consent before providing basic demographic details on Qualtrics. Prior to the reasoning task, participants were given written and verbal instructions about the general procedure including a brief explanation of the eye-tracking equipment. The chair and chinrest were then adjusted to a comfortable height for the participant and the eye-tracking procedure began. A nine-point calibration was conducted, and three practice items were completed. Participants were then offered the chance to ask questions and adjust their position. If needed, the calibration was completed again.

Three 3-min breaks were included in which participants were advised to take their chin off the rest and close their eyes, if they felt comfortable doing so. This was to help hydrate the eyes and reduce blinking during the trials, and to lower fatigue effects. A nine-point calibration was conducted after each break. In case the participant moved during a trial block, a one-point calibration was presented prior to each item. If this was failed, the nine-point calibration was conducted again before continuing through the remaining items in the block.

Results

Performance

As expected, performance for lure items increased with test block; specifically, performance improved from T1 to T2 but remained stable from T2 to T4 (see Fig. 5). In line with our

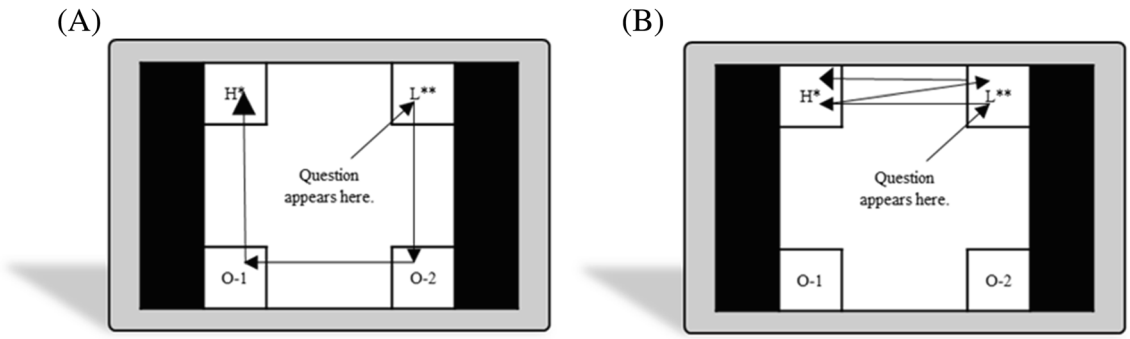

Fig. 4 Hypothetical gaze patterns reflecting uncertainty on heuristic lure trials that stems from (**A**) non-specific or (**B**) process-specific phenomena. Arrows represent gaze movements starting at the question and moving to the final response selection. The size of the arrowhead represents the number of fixations and amount of dwell. *H* heuristic response option, *L* logical response option, *O-1 and O-2* other incorrect response options. * Indicates the selected option, ** indicates the other-relevant option

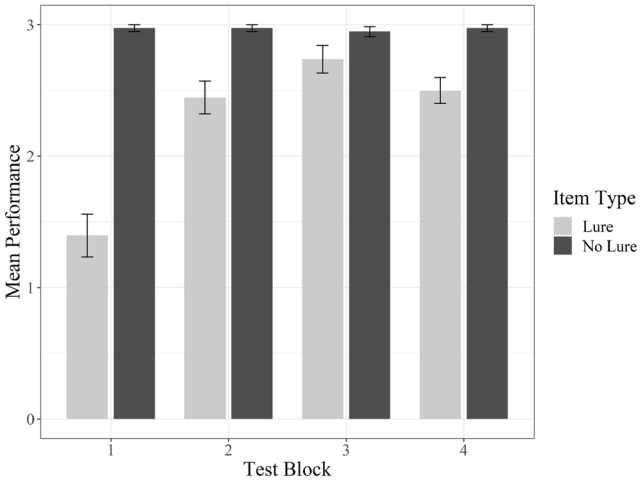

Fig. 5. Performance by test block and item type. Error bars ± 1 SE

hypotheses, performance for no-lure items remained high and stable across test blocks (see Fig. 5). These contrasting results for lure and no-lure items indicated that the no-lure items were suitable baseline trials for examining whether, in comparison to the correct-no-lure items, uncertainty was registered for heuristic lure trials (see *Uncertainty*).

A repeated-measures ANOVA was used to formally examine the effects of item type (lure and no lure) and test block (T1, T2, T3, T4) on performance (scores could range from 0 to 3). Where sphericity was violated, Greenhouse-Geisser adjusted results are reported. Results of the ANOVA are reported in Table 1.

As reported in Table 1, main effects of item type and test block were observed, however, these were qualified by a significant two-way interaction between item type and test block (see Fig. 5). To explore this further we examined the effect of test block for no-lure and lure items separately. Test block did not have a significant effect on performance on no-lure items, $F(3, 103.10) = .20$, $p = .899$, $\eta^2_p = .005$. However, it did have a significant effect on performance on lure items, $F(2.29,84.78) = 43.35$, $p < .001$, $\eta^2_p = .540$. Pairwise comparisons between test blocks were examined against a Bonferroni adjusted alpha of .017. For lure items, performance significantly increased from T1 ($M = 1.40$, $SE = .16$) to T2 ($M = 2.45$, $SE = .12$), $F(1,37) = 59.95$, $p < .001$, $\eta^2_p = .618$, and from T2 to T3 ($M = 2.74$, $SE = .11$), $F(1,37) = 8.52$, $p = .006$, $\eta^2_p = .187$. However, there was no significant difference between performance at T3 and T4 ($M = 2.50$, $SE = .10$), $F(1,37) = 4.64$, $p = .037$, $\eta^2_p = .111$.

Table 1. ANOVA with effects of test block and item type on performance

Source	Df_{Source}, Df_{Error}	F	n^2_p	p
Item Type	1, 37	54.08	.594	<.001
Test Block	19.84, 20.41	35.74	.493	<.001
Item Type * Test Block	2.35, 86.85	42.77	.536	<.001

Cognitive pathways

To examine whether the sequence of cognitive events leading to the respondents' solutions were more indicative of corrective or intuitive cognitive pathways, we compared confidence and gaze-patterns for lure items with correct responses (correct lure trials) and lure items with heuristic responses (heuristic lure trials). For these two trial types, we assessed the number of fixations and amount of dwell that occurred with each AOI assigned to a response option. Trials on which participants gave neither the correct nor heuristic response were removed for this analysis. Recall that selection-contingent coding was applied to the multiple-choice options such that they were labelled as the 'Selected' AOI (correct for correct trials and heuristic for heuristic trials), the 'Other-Relevant' AOI (heuristic for correct trials and correct for heuristic trials), and the 'Other-1' and 'Other-2' AOI (corresponding to the remaining two non-selected options).

If our data indicated the use of corrective pathways, we expected to observe a difference in the pattern of fixations and dwell upon the four AOIs for correct versus heuristic lure trials. In our analysis this is captured by the two-way interaction between AOI and trial (correct lure, heuristic lure). Specifically, we would expect to observe greater consideration of the heuristic option than the other two foil options on correct lure trials compared to heuristic lure trials. However, we observed a similar pattern of fixations across the four response options for both correct and heuristic trials (Fig. 6). Notably, for correct trials the proportion of fixations upon the heuristic response option was no greater than the proportion of fixations on the two foil options.

To analyse the cognitive pathways we used a linear mixed model with AOI (Selected, Other-Relevant, Other-1, Other-2) nested within trial (correct, heuristic), trial within item[6] (1,2,3), item within test block (T1,T2,T3,T4), and test block within participant. The model included four predictors: item, test block, trial, and AOI. The dependent variable was the proportion of fixations. There was a significant main effect of AOI, $F(3,1190.07) = 149.95$, $p < .001$. Therefore, pairwise comparisons were conducted to examine the main effect of AOI more closely. Tests were compared to a Bonferroni adjusted alpha of .008. The proportion of fixations was greater for the 'Selected' AOI than all other AOIs. The proportion of

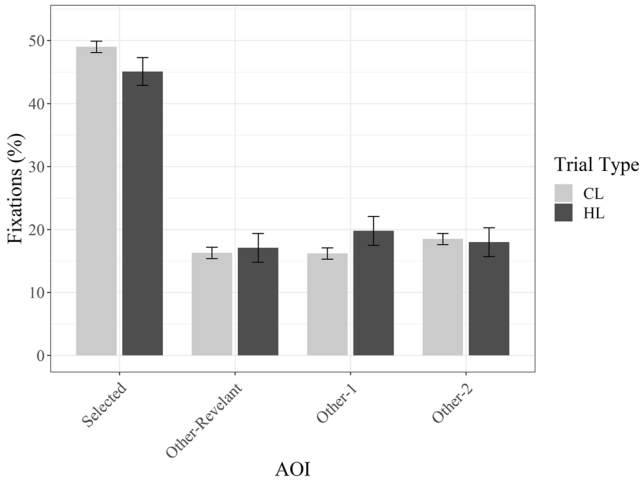

Fig. 6. Proportion of fixations on each area of interest (AOI) for correct and heuristic responding on lure items. *CL* correct lure trials, *HL* heuristic lure trials. Error bars ± 1 SE

fixations was not significantly different between any of the three 'Other-' AOIs (Table 2).

The two-way interaction between AOI and trial was not significant, $F(3,1193.79) = 1.765$, $p = .152$. This indicated that the pattern of fixations did not differ between correct- and heuristic-trials. Specifically, there was no evidence that correct responding was associated with a greater consideration of the heuristic response than heuristic responding was of the correct response. We note, additionally, that the proportion of fixations on the heuristic option did not show a trend in line with the corrective pathway hypothesis.

In sum, we found no evidence for the use of corrective cognitive pathways rather, the results lend support to the interpretation of correct responding occurring via intuitive cognitive pathways. Gaze patterns showed no evidence for the consideration or override of the heuristic response. Rather, correct participants appeared to primarily consider the correct response and gave no more consideration to the heuristic option than the other foil options. Additionally, there was no interaction with test block, which indicates that there was no evidence for corrective pathways prior to training at Test Block 1, or as training increased at Test Blocks 2 to 4 (see Appendix 2, Table 6). The same analysis was conducted for dwell and the interpretation of the results did not change (see Appendix 2, Table 8).

Table 2. Pairwise comparisons of fixations by area of interest (AOI)

Comparison	Mean difference	SE	df	F	p
1 - 2	0.304	.019	896.74	256.00	<.001
1 - 3	0.291	.017	1635.66	293.01	<.001
1 - 4	0.288	.017	1410.63	287.00	<.001
2 - 3	-0.013	.019	833.51	0.47	0.478
2 - 4	-0.016	.017	1656.80	0.89	0.346
3 - 4	-0.002	.019	830.17	0.01	0.905

The comparisons are coded for each AOI such that 1 = 'Selected', 2 = 'Other Relevant', 3 = 'Other-1', 4 = 'Other-2'

Uncertainty

To examine uncertainty experienced by reasoners completing the CRT, we compared uncertainty for lure trials on which heuristic responses were made (heuristic lure trials) and uncertainty for no lure trials on which correct responses were made (correct-no lure trials). As expected, we observed greater uncertainty for heuristic lure trials than correct-no lure trials. Specifically, we found that participants' reverse-coded confidence ratings and consideration of non-selected multiple-choice options were higher for heuristic lure trials than correct-no lure trials. For the interested reader, an additional exploratory comparison of confidence- and gaze-based uncertainty on correct lure trials and heuristic lure trials is provided in Appendix 3.

To examine the relationships between the uncertainty measures, we used a confidence-based uncertainty index and two novel gaze-based uncertainty indices. The

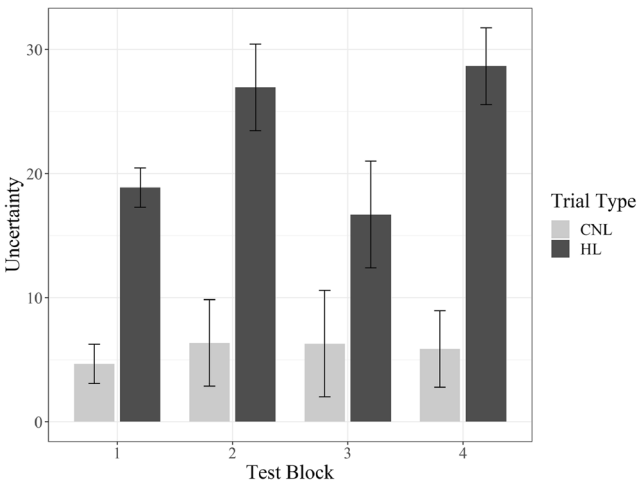

Fig. 7. Confidence-based uncertainty (reverse coded confidence out of 100) by test block and trial type. *CNL* correct-no lure trials, *HL* heuristic lure trials. Error bars ± 1 SE

confidence-based index was calculated by reverse-coding participants' confidence ratings, and gaze-based indices were generated as the proportion of fixations and dwell on the participants' non-selected responses for each trial. We ran three bivariate correlations between these indices; the confidence-based index was weakly correlated with the fixations-based index (r = .133, p < .001) and the dwell-based index (r = .185, p < .001). The fixation- and dwell-based indices were strongly correlated with each other (r = .874, p < .001). This suggests weak convergent validity between the confidence-based and gaze-based measures and, as one would expect, strong convergent validity for the two gaze-based measures.

Confidence-based uncertainty

As expected, confidence-based uncertainty was greater for heuristic lure trials than correct-no lure trials (see Fig. 7). A linear mixed model was used to formally examine uncertainty measured via reverse-coded confidence ratings. The model included the predictors: item (1,2,3), test block (T1, T2, T3, T4) and trial type (heuristic lure, correct-no lure). Item did not have a significant effect, $F(2,525) = 1.21$, $p = .298$. The main effect of trial type was significant, $F(1,525) = 103.51$, $p < .001$, as was the main effect of test block , $F(3,525) = 3.10$, $p = .026$. These were qualified by a two-way interaction between test block and trial type, $F(3,525) = 3.01$, $p = .026$. The simple effects were examined for trial type at each level of test block. The results were compared to a Bonferroni-adjusted alpha of .0125. Confidence-based uncertainty was significantly greater on heuristic lure than correct-no lure trials at T1, $F(1,525) = 77.54$, $p < .001$; T2, $F(1,525) = 33.76$, $p < .001$; T4, $F(1,525) = 52.83$, $p < .001$; and marginally significant at T3, $F(1,525) = 5.62$, $p = .018$.

Gaze-based uncertainty

As expected, gaze-based uncertainty – indicated by the greater consideration of the non-selected responses – was higher for heuristic lure trials than correct-no lure trials (Fig. 8). When responding heuristically on lure problems, participants visually considered the non-selected options to a greater extent than when responding correctly on the no-lure problems. Additionally, gaze-patterns indicate that this uncertainty reflects a non-specific rather than process-specific signal. That is, the increased consideration of non-selected options for heuristic lure trials did not coincide with the particular consideration of the correct option over the other foil options.

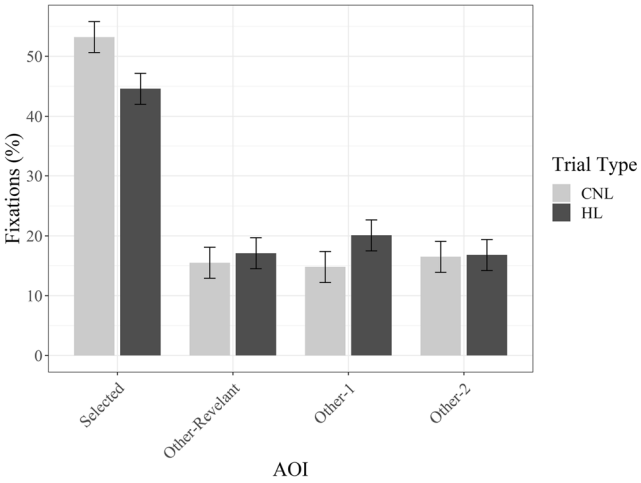

Fig. 8. Fixations by area of interest (AOI) and trial type. *CNL* correct-no lure, *HL* heuristic lure. Error bars ± 1 SE

A linear mixed model was used to examine uncertainty via fixations. The dependent variable was the proportion of fixations that occurred within an AOI out of the total fixations that occurred within all four AOIs. The predictors were item (1,2,3), test block (T1, T2, T3, T4), trial type (heuristic lure, correct-no lure), and AOI. The model is presented in Table 3.

The significant two-way interaction between trial type and AOI was examined by comparing the proportion of fixations for heuristic lure and correct-no lure trial types for each AOI. The analysis was compared to a Bonferroni adjusted alpha of .0125 (see Table 4).

The significantly higher proportion of fixations on the selected response for the correct-no lure trials than the

Table 3. Linear mixed model for gaze-based uncertainty measured as the proportion of fixations on each area of interest (AOI)

	F	Df	*p*
Item	.04	2	.962
Test Block	.01	3	.993
Trial Type	.09	1	.767
AOI	152.14	3	<.001
Test Block * AOI	.82	9	.598
Test Block * Trial Type	.03	3	.993
Trial Type * AOI	5.55	9	.001
Test Block * Trial Type * AOI	1.21	9	.286

Df error = 2098

Table 4. Pairwise comparisons for the proportions of fixations for heuristic lure (HL) and correct-no lure (CNL) trials by area of interest (AOI)

AOI (HL/CNL)	F	*p*
Selected/Selected	11.38	.001
Other-Relevant/Other-1	0.42	.518
Other-1/Other-2	3.86	.050
Other-2/Other-3	.04	.841

df Source = 1; df Error = 2098

heuristic lure trials indicates that participants examined the non-selected responses to a greater extent on heuristic lure trials (see Selected/Selected, Table 4). The mean proportions and standard errors are presented in Fig. 8. These results suggest greater information search and uncertainty for heuristic lure trials than correct-no lure trials as indicated via gaze patterns.

Additionally, there was no evidence that the correct option was considered to a greater extent than the remaining foil options on heuristic trials (Fig. 8). This suggests that uncertainty was more likely to stem from a non-specific phenomenon when completing the task rather than from the process-specific phenomenon reflecting conflict between heuristic and logical solution processes. The same analysis was conducted using dwell as the dependent variable and the interpretation of the results did not change (see Appendix 2, Tables 9 and 10).

Discussion

The current study employed a training and eye-tracking paradigm to explore the cognitive pathways and uncertainty experienced by reasoners solving the CRT. We found that performance on the CRT-like lure items improved with training, while performance on no-lure items remained consistently high. Across training, there was greater evidence for the use of intuitive than corrective pathways being used to reach correct solutions; these findings supported the hybrid models' interpretation of reasoning on the CRT over the default-intervention interpretation. We also observed greater uncertainty for participants providing heuristic responses on lure items than correct responses on no-lure items. Gaze-patterns indicated that this uncertainty was more likely to have stemmed from a non-specific signal than a process-specific signal.

Eye-tracking measures were employed to compare the cognitive pathways (corrective or intuitive) used on correct and heuristic trials on lure items. The analysis of both fixations and dwell revealed that respondents looked at the response they eventually selected more often and for longer than they looked at the alternatives. The default-intervention interpretation of correct responding on the CRT stipulates that an initial incorrect response is generated via Type 1 processes but that this response can be overridden by Type 2 processes that can then generate the correct response (Kahneman & Frederick, 2005). This theory would be supported if participants had visually considered the heuristic response to a greater extent than the other non-selected alternatives. However, there was no evidence to suggest that correct respondents considered the heuristic option to a greater extent than the other incorrect responses. Rather, participants seem to follow an intuitive cognitive pathway, making their logical choice without first considering the heuristic option.

It is worth noting again that this analysis examined the proportion of fixations and dwell on the four multiple choice options, not the order in which the responses were considered. The assessment of order effects would be difficult with eye-tracking measures; for example, participants may conduct a visual scan of the response options before considering certain options in more depth. However, the default intervention hypothesis – that the heuristic solution is considered by default and then potentially overridden by a logical solution – can be examined by assessing one of its implications. In particular, the implication that the heuristic option should be considered to a greater extent than the other non-heuristic foil options. Counter to the corrective, default-intervention interpretation of correct responding on the CRT, we did not find any evidence to support this assertion.

Previous studies have shown varied results in relation to cognitive pathways. In line with the default-intervention interpretation, Travers et al.'s (2016) mouse-tracking study found that correct responders were "attracted" to the heuristic response more than heuristic responders were "attracted" to the correct response. In contrast, Szaszi et al.'s (2017; see also Mata et al., 2013, Study 5) think-aloud study found that most correct respondents began reasoning in line with the correct, not heuristic, solution. The current study substantiates Szaszi et al.'s (2017) results, suggesting that correct responding does not follow a default-intervention, corrective cognitive pathway. Rather, it appears that correct responding is more easily explained by hybrid dual process models that emphasise the ability of reasoners to reach logical solutions without needing to override an initial heuristic incorrect response.

The current study observed greater confidence- and gaze-based uncertainty on heuristic lure trials than correct-no lure trials. These findings differ from Travers et al.'s (2016) study of mouse movements that indicated lower uncertainty was experienced on heuristic lure than correct-no lure trials. Together with the current findings, it is evident that the tools used to measure uncertainty on the CRT can generate different results, possibly because they tap into distinct representations of this psychological phenomenon. The indicators of uncertainty may differ as a factor of awareness; for instance, think-aloud protocols may reflect uncertainty signals present at a higher level of awareness than mouse-tracking protocols, and mouse-tracking protocols may reflect uncertainty signals at a higher level of awareness than eye-tracking protocols. We note that even in the current study, the new gaze-based measures of uncertainty were only weakly associated with the confidence-based measure. Future comparisons of techniques for assessing uncertainty may aid our understanding of the psychological phenomena underpinning these effects, and of cognitive uncertainty more broadly.

The use of eye tracking also allowed for the examination of the specificity of uncertainty. The gaze patterns suggested that uncertainty stemmed from a non-specific effect rather than a process-specific effect (i.e., the specific competition between the heuristic and logical processes). That is, for trials on which the participant had selected the incorrect heuristic response, the proportion of fixations and dwell on the non-selected correct logical option was no greater than the proportion of fixations and dwell on the other non-selected options. This suggests that as the participants engaged in greater information search and looked at the alternative responses, they were not considering the logical response any more than the other foils. Hence, there was no evidence that uncertainty experienced by reasoners giving the heuristic response was due to the competition between cognitive processes that, if completed, would lead to heuristic versus logical responses.

The empirical specification of uncertainty, both in the current article and in previous studies (Bago et al., 2019; Travers et al., 2016), offers a strong foundation for uncertainty to be examined in greater depth in the context of thinking and reasoning. More research is needed to understand the underlying processes and architecture of uncertainty so that it can be incorporated more effectively in models of reasoning, in particular, to explain its role in the engagement of effortful thinking – a primary enquiry for dual process theorists. This research may benefit from the surrounding areas of investigation. For example, philosophers of mind and neuroscientists are investigating the role of prediction error in the accrual of mental resources and the propagation of mental activity (e.g., Clark, 2016; Parr & Friston, 2017) and metacognitive scholars examining the relationships between individual differences, confidence, and uncertainty (e.g., Jackson et al., 2016; Stankov et al., 2015).

We included a training manipulation in our paradigm to increase within-subject variance in accuracy for a more rigorous examination of cognitive pathways. However, it can also be used to examine the potential effects of training on pathways and uncertainty. While Purcell et al. (2021; see also Stanovich, 2018) postulated that uncertainty and corrective reasoning on the CRT may be more likely at intermediate stages of domain-specific experience, there is no previous empirical support for these claims. In the current study, training had no effect on cognitive pathways or uncertainty. The findings at Test Block 1, prior to training, did not differ from the remaining post-training test blocks; however, there could have been some systematic variation that, while undetected in the current study, are worth investigating more directly in future experiments. In particular, it would be interesting to explore whether training effects on pathways and uncertainty emerge when individual differences, such as intelligence, working memory capacity, or pre-existing numeracy, are considered.

This work examined cognitive pathways and uncertainty on the frequently used CRT-like tasks. Using rigorous gaze-based measures of cognitive pathways, we found greater evidence for the hybrid dual process models than the default intervention models, by demonstrating that correct responding was more likely to be associated with intuitive than corrective cognitive pathways. We found support for the assertion that people giving heuristic responses on the CRT register this inaccuracy at some level, and that this registration can be captured by both confidence- and gaze-based measures of uncertainty. The gaze patterns suggest this uncertainty reflected a non-specific signal rather than a process-specific signal. The current article presents new techniques and empirical findings for the investigation of cognitive pathways and uncertainty for reasoning on the CRT. These findings have critical implications for the theoretical development of cognitive pathways and uncertainty within dual process models of reasoning.

Open Practices Statement The data for all experiments are available. None of the experiments were preregistered.

Appendix 1

Table 5. Order of item presentation by counterbalance condition

Counter-balance	Block	Block description	Item*	Item description
1	1	Practice	1	Practice
			2	Practice
			3	Practice
	2	Test Block 1	4	Test Item – Lure
			5	Test Item – Lure
			6	Test Item – Lure
			7	Test Item – No Lure
			8	Test Item – No Lure
			9	Test Item – No Lure
	3	Training Block 1	28	Training Item – Lure
			29	Training Item – Lure
			30	Training Item – Lure
			31	Training Item – No Lure
			32	Training Item – No Lure
			33	Training Item – No Lure
	4	Test Block 2	10	Test Item – Lure
			11	Test Item – Lure
			12	Test Item – Lure
			13	Test Item – No Lure
			14	Test Item – No Lure
			15	Test Item – No Lure
	5	Training Block 2	34	Training Item – Lure
			35	Training Item – Lure
			36	Training Item – Lure
			37	Training Item – No Lure
			38	Training Item – No Lure
			39	Training Item – No Lure
	6	Test Block 3	16	Test Item – Lure
			17	Test Item – Lure
			18	Test Item – Lure
			19	Test Item – No Lure
			20	Test Item – No Lure
			21	Test Item – No Lure
	7	Training Block 3	40	Training Item – Lure
			41	Training Item – Lure
			42	Training Item – Lure
			43	Training Item – No Lure
			44	Training Item – No Lure
			45	Training Item – No Lure
	8	Test Block 4	22	Test Item – Lure
			23	Test Item – Lure
			24	Test Item – Lure
			25	Test Item – No Lure
			26	Test Item – No Lure
			27	Test Item – No Lure

Table 5. (continued)

Counter-balance	Block	Block description	Item*	Item description
2	1	Practice	1	Practice
			2	Practice
			3	Practice
	2	Test Block 1	45	Test Item – Lure
			44	Test Item – Lure
			43	Test Item – Lure
			42	Test Item – No Lure
			41	Test Item – No Lure
			40	Test Item – No Lure
	3	Training Block 1	39	Training Item – Lure
			38	Training Item – Lure
			37	Training Item – Lure
			36	Training Item – No Lure
			35	Training Item – No Lure
			34	Training Item – No Lure
	4	Test Block 2	33	Test Item – Lure
			32	Test Item – Lure
			31	Test Item – Lure
			30	Test Item – No Lure
			29	Test Item – No Lure
			28	Test Item – No Lure
	5	Training Block 2	27	Training Item – Lure
			26	Training Item – Lure
			25	Training Item – Lure
			24	Training Item – No Lure
			23	Training Item – No Lure
			22	Training Item – No Lure
	6	Test Block 3	21	Test Item – Lure
			20	Test Item – Lure
			19	Test Item – Lure
			18	Test Item – No Lure
			17	Test Item – No Lure
			16	Test Item – No Lure
	7	Training Block 3	15	Training Item – Lure
			14	Training Item – Lure
			13	Training Item – Lure
			12	Training Item – No Lure
			11	Training Item – No Lure
			10	Training Item – No Lure
	8	Test Block 4	9	Test Item – Lure
			8	Test Item – Lure
			7	Test Item – Lure
			6	Test Item – No Lure
			5	Test Item – No Lure
			4	Test Item – No Lure

See https://osf.io/ej3n2/ for full items. *The order of items was randomised within block

Appendix 2

Table 6. Results for the linear mixed model used to examine cognitive pathways (via fixations)

	F	df Source	df Error	*p*
Item	.013	2	698.82	.987
Test Block	.003	3	800.15	>.999
Trial	.000	1	803.98	.990
AOI	149.95	3	1190.07	<.001
Test Block* Trial	.000	3	782.61	>.999
Test Block* AOI	.942	9	1193.56	.487
Trial* AOI	1.765	3	1193.79	.152
Test Block* Trial* AOI	1.155	9	1192.58	.321

Table 7 Results for the linear mixed model used to examine cognitive pathways (via dwell)

	F	df Source	df Error	*p*
Item	.007	2	671.10	.993
Test Block	.002	3	766.97	>.999
Accuracy	.002	1	771.13	.966
AOI	407.295	3	1170.21	<.001
Test Block* Accuracy	.001	3	749.82	>.999
Test Block* AOI	1.280	9	1173.31	.243
Accuracy* AOI	1.826	3	1173.52	.141
Test Block* Accuracy* AOI	1.31	9	1172.45	.216

Table 8. Pairwise comparisons for cognitive pathways (via dwell)

Comparison	Mean difference	SE	df Error	F	*p*
1 - 2	0.447	.017	876.16	693.00	<.001
1 - 3	0.442	.015	1636.99	861.62	<.001
1 - 4	0.428	.016	1422.38	751.92	<.001
2 - 3	-0.005	.017	815.69	0.08	0.771
2 - 4	-0.019	.015	1657.72	1.55	0.231
3 - 4	-0.014	.017	813.01	0.66	0.417

The comparisons are coded for each AOI such that 1.00 = 'Selected', 2.00 = 'Other Relevant', 3.00 = 'Other-1', 4.00 = 'Other-2'

Table 9. Results for the linear mixed model used to examine uncertainty (via dwell)

	F	df Source	*p*
Item	.05	2	.949
Test Block	.01	3	.998
Problem Type	.02	1	.876
AOI	409.74	3	<.001
Test Block* AOI	1.57	9	.119
Test Block* Problem Type	.01	3	.999
Problem Type * AOI	6.34	9	<.001
Test Block* Problem Type * AOI	.861	9	.560

Df error = 2098

Table 10. Pairwise comparisons for uncertainty (via dwell) for heuristic lure (HL) and correct-no lure (CNL) trials

AOI (HL/CNL)	F	*p*
Selected/Selected	13.93	<.001
Other-Relevant/Other-1	0.82	.366
Other-1/Other-2	1.69	.193
Other-2/Other-3	1.48	.224

df Error = 2098

Appendix 3

A linear mixed model was used to examine uncertainty measured via reverse-coded confidence ratings for heuristic lure and correct lure trials. The model included the predictors: item (1,2,3), test block (T1, T2, T3, T4) and trial type (heuristic lure, correct lure). The effects are reported in Table 11. Trial type had a main effect, such that uncertainty was higher on heuristic lure trials (M = 21.65, SD = 2.27) than correct lure trials (M = 10.46, SD = 1.10). Uncertainty for item 1(M = 16.57, SD = 1.75) and item 2 (M = 18.11, SD = 1.63) was not significantly different, $F(1,419) = .092$, $p = .401$, $\eta^2_p = .002$; nor was it significantly different for item 1 and item 3 (M = 13.46, SD = 1.60), $F(1,419) = 3.57$, $p = .060$, $\eta^2_p = .008$; but it was for significantly higher for item 2 than item 3, $F(1,419) = 7.99$, $p = .005$, $\eta^2_p = .019$;

Table 11. Results for the linear mixed model used to examine confidence-based uncertainty

	F	df Source	df Error	*p*
Item	4.15	2	419	.017
Test Block	.62	3	419	.605
Trial Type	23.50	1	419	<.001
Test Block * Trial Type	.87	3	419	.455

A linear mixed model was used to examine gaze-based uncertainty measured via fixations by trial type (heuristic lure and correct lure trials). These effects are reported in Table 12. There was no interaction between trial type and AOI, which indicates that gaze-based uncertainty did not differ between heuristic and correct trials for lure items. AOI had a main effect which indicated that the number of fixations occurring on the response that the participants finally selected was greater than for the remaining three options (see Tables 12 and 13). There was no difference between the number of fixations on the non-selected responses (Table 13).

Table 12. Results for the linear mixed model used to gaze-based uncertainty

	F	df Source	*P*
Test Block	.001	3	>.999
Item	.014	2	.986
Trial Type	.006	1	.938
AOI	152.87	3	<.001
Test Block* AOI	.945	9	.485
Test Block * Trial Type	.003	3	>.999
Trial Type * AOI	1.687	3	.168
Test Block * Trial Type * AOI	1.112	9	.351

Df error = 1678

Table 13. Pairwise comparisons for gaze-based uncertainty by area of interest (AOI)

AOI	F	p
Selected - Other-Relevant	258.72	<.001
Selected - Other-1	332.45	<.001
Selected - Other-2	291.07	<.001
Other-Relevant - Other-1	0.53	.463
Other-Relevant - Other-2	0.84	.348
Other-1 - Other-2	0.01	.928

df Error = 1678. Selected = correct response for correct lure trials and heuristic response for heuristic lure trials. Other-relevant = heuristic response for correct trials and correct response for heuristic trials. Other-1 and Other-2 AOIs were randomly assigned to the remaining foil options

Acknowledgements Zoe A. Purcell acknowledges support from the Institute for Advanced Study in Toulouse and the grant ANR-17-EURE-0010 Investissements d'Avenir.

References

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

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

Bago, B., Raoelison, M., & De Neys, W. (2019). Second-guess: Testing the specificity of error detection in the bat-and-ball problem. *Acta Psychologica*, *193*, 214–228. https://doi.org/10.1016/j.actpsy.2019.01.008

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?. Quarterly Journal of Experimental Psychology A Human Experimental Psychology, *56*(6), 1053–1077. https://doi.org/10.1080/02724980244000729

Ball, L. J., Phillips, P., Wade, C. N., & Quayle, J. D. (2006). Effects of belief and logic on syllogistic reasoning: Eye-movement evidence for selective processing models. *Experimental Psychology*, *53*(1), 77–86. https://doi.org/10.1027/1618-3169.53.1.77

Bruckmaier, G., Binder, K., Krauss, S., & Kufner, H.-M. (2019). An eye-tracking study of statistical reasoning with tree diagrams and 2 × 2 tables. *Frontiers in Psychology*, *10*, 632. https://doi.org/10.3389/fpsyg.2019.00632

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Crutcher, R. J. (1994). Telling what we know: The use of verbal report methodologies in psychological research. *Psychological Science*, *5*(5), 241–241. https://doi.org/10.1111/j.1467-9280.1994.tb00619.x

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

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

De Neys, W. (2018). *Dual process theory 2.0*. Routledge. https://doi.org/10.4324/9781315204550

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

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

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

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

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

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

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

Glöckner, A., & Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1055–1075. https://doi.org/10.1037/0278-7393.34.5.1055

Glöckner, A., & Herbold, A.-K. (2011). An eye-tracking study on information processing in risky decisions: Evidence for compensatory

strategies based on automatic processes. *Journal of Behavioral Decision Making*, *24*(1), 71–98. https://doi.org/10.1002/bdm.684

Green, H. J., Lemaire, P., & Dufau, S. (2007). Eye movement correlates of younger and older adults' strategies for complex addition. *Acta Psychologica*, *125*(3), 257–278. https://doi.org/10.1016/J.ACTPSY.2006.08.001

Hoover, J. D., & Healy, A. F. (2019). The bat-and-ball problem: Stronger evidence in support of a conscious error process. *Decision*, *6*(4), 369–380. https://doi.org/10.1037/dec0000107

Jackson, S. A., Kleitman, S., Stankov, L., & Howie, P. (2016). Decision pattern analysis as a general framework for studying individual differences in decision making. *Journal of Behavioral Decision Making*, *29*(4), 392–408. https://doi.org/10.1002/bdm.1887

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

Kahneman, D., & Frederick, S. (2005). A Model of Heuristic Judgment. In K. Holyoak & R. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 267–293). Cambridge University Press.

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology*, *105*(3), 353–373. https://doi.org/10.1037/a0033640

Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Scientific Reports*, *7*(1), 1–21. https://doi.org/10.1038/s41598-017-15249-0

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

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgments and decisions. *Current Directions in Psychological Science*, *21*(1), 31–35. https://doi.org/10.1177/0963721411429960

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying Item Response Theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453–469. https://doi.org/10.1002/bdm.1883

Purcell, Z. A., Wastell, C. A., & Sweller, N. (2021). Domain-specific experience and dual-process thinking. *Thinking and Reasoning*, *27*(2), 239–267. https://doi.org/10.1080/13546783.2020.1793813

Raoelison, M. T. S., Thompson, V. A., & De Neys, W. (2020). The smart intuitor: Cognitive capacity predicts intuitive rather than deliberate thinking. *Cognition*, *204*, 104381. https://doi.org/10.1016/j.cognition.2020.104381

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Stankov, L., Kleitman, S., & Jackson, S. A. (2015). Measures of the Trait of Confidence. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of Personality and Social Psychological Constructs* (pp. 158–189). Academic Press. https://doi.org/10.1016/B978-0-12-386915-9.00007-3

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking and Reasoning*, *24*(4), 423–444. https://doi.org/10.1080/13546783.2018.1459314

Stephen, D. G., Boncoddo, R. A., Magnuson, J. S., & Dixon, J. A. (2009). The dynamics of insight: Mathematical discovery as a phase transition. *Memory & Cognition*, *37*(8), 1132–1149. https://doi.org/10.3758/MC.37.8.1132

Stupple, E. J. N., Pitchford, M., Ball, L. J., Hunt, T. E., Steel, R., & Antonietti, A. (2017). Slower is not always better: Response-time evidence clarifies the limited role of miserly information processing in the Cognitive Reflection Test. *PloS One.*, *12*(11), e0186404. https://doi.org/10.1371/journal.pone.0186404

Szaszi, B., Palfi, B., Szollosi, A., Kieslich, P. J., & Aczel, B. (2018). Thinking dynamics and individual differences: Mouse-tracking analysis of the denominator neglect task. *Judgment & Decision Making*, *13*(1), 23–32.

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: Exploring the ways individuals solve the test. *Thinking and Reasoning*, *23*(3), 207–234. https://doi.org/10.1080/13546783.2017.1292954

Thompson, V. A. (2021). Eye-tracking IQ: Cognitive capacity and strategy use on a ratio-bias task. *Cognition*, *208*, 104523–104523. https://doi.org/10.1016/j.cognition.2020.104523

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking and Reasoning*, *20*(2), 215–244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., & Morsanyi, K. (2012). Analytic thinking: Do you feel like it? *Mind and Society*, *11*(1), 93–105. https://doi.org/10.1007/s11299-012-0100-6

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. *Cognition*, *150*, 109–118. https://doi.org/10.1016/j.cognition.2016.01.015

Velichkovsky, B. M. (2014). From levels of processing to stratification of cognition: Converging evidence from three domains of research. In B. H. Challis & B. M. Velichkovsky (Eds.), *Stratification in cognition and consciousness* (p. 203). J. Benjamins. https://doi.org/10.1075/aicr.15.13vel

Velichkovsky, B. M., Rothert, A., Kopf, M., Dornhöfer, S. M., & Joos, M. (2002). Towards an express-diagnostics for level of processing and hazard perception. *Transportation Research.*, *5*(2), 145–156. https://doi.org/10.1016/S1369-8478(02)00013-X