Eye Movements Reveal that Low Confidence Precedes Deliberation

Zoe A. Purcell1,2,\*, Colin A. Wastell1, and Naomi Sweller1

1. Department of Psychology, Macquarie University, Sydney, Australia
2. Artificial and Natural Intelligence Toulouse Institute (ANITI), University of Toulouse, Toulouse, France

\*Corresponding author: zoe.purcell@iast.fr

Acknowledgements: Associate Professor Wastell passed away since the initial submission of this manuscript and this paper is published in his memory. Zoe A. Purcell acknowledges support from the Institute for Advanced Study in Toulouse and the grant ANR-17-EURE-0010 Investissements d’Avenir.

Abstract

 Contemporary dual process models of reasoning maintain that there are two types of thinking – intuitive and deliberative – and that low confidence predicts the engagement of deliberation. Previous studies examining the confidence-deliberation relationship have been limited by 1) issues of endogeneity and between-subject comparisons – concerns that we address by employing debias training; and 2) measures of confidence that are taken relatively late in the reasoning process – a concern that we address by employing a real-time eye-tracking measure of confidence. Self-reported and eye-tracked confidence were negatively related to deliberative thinking, providing novel evidence for the time-course of the confidence-deliberation relationship, and revealing that lowered confidence precedes deliberation.

*Key words*: Dual process, Reasoning, Uncertainty, Confidence, CRT.

Eye Movements Reveal that Low Confidence Precedes Deliberation

Dual process theories distinguish between intuitive and deliberative thinking (e.g., Evans & Stanovich, 2013; Kahneman, 2011). People are thought to rely heavily on intuitive thinking – preserving limited deliberative resources where possible (Kahneman, 2011; Stanovich, 2009). Therefore, a fundamental question for reasoning scholars is *when* and *how* people engage in deliberative thinking. Contemporary reasoning models propose that metacognitive factors related to confidence and uncertainty[[1]](#footnote-2) such as “feelings of rightness” or “conflict” may be involved (e.g., De Neys, 2012; Pennycook, Fugelsang, & Koehler, 2015; Thompson & Morsanyi, 2012; for a review De Neys, 2018).

To elaborate, we will explore three models of reasoning that make explicit hypotheses about the relationship between metacognitions and deliberation. The logical intuition model proposed that when we encounter a problem or situation multiple cognitive process can be triggered; the competition between these processes generates metacognitive “conflict” which, in turn, triggers the engagement of deliberation (De Neys, 2012; De Neys, 2014). Similarly, the three-stage dual-process model also referred to conflict as a determinant of deliberation but explicitly separated an initial stage in which Type 1 processes are triggered, the intermediary stage of conflict detection, and the final stage of analytic thinking (Pennycook et al., 2015). This allowed for the distinction of reasoning failures as resulting from poor conflict detection or poor analytic thinking.

Additionally, the metacognitive model suggested that the slower or more difficult it is to generate a solution, the lower the reasoner’s metacognitive “feeling of rightness” and, if low enough, this metacognition triggers deliberation (Thompson, Prowse Turner, & Pennycook, 2011). Although they emphasise different aspects of metacognitions, these three models converge on the idea that lower confidence (higher *conflict* or lower *feeling of rightness*) precedes the engagement of deliberative thinking.

Studying the relationship between confidence and deliberation requires combining effective methods for measuring these two phenomena. Effective measures for both confidence and deliberation have been developed for use with the CRT – a well-known task that commonly elicits errors despite the respondents usually having the sufficient mindware to reach the correct solution (Frederick, 2005). The most well-known CRT item states, “A bat and a ball cost $1.10 together. The bat costs $1 more than the ball. How much does the ball cost?” The correct answer is 5 cents; however, the most common response is 10 cents.

Confidence and reasoning have been successfully examined in studies using the CRT. Such studies largely employed self-reported confidence ratings (De Neys et al., 2013; Hoover & Healy, 2019; Purcell et al., 2022) and self-reported feelings of rightness (Gangemi, Bourgeois-Gironde & Mancini, 2015). However, a recent study demonstrated that eye-tracking can be used to assess confidence on the CRT at the time of reasoning and without depending on self-report (Purcell et al., 2022). Crucially, this technique allows confidence tracking from the outset of reasoning which can be used to test the contemporary models’ proposals that lower confidence *precedes* deliberation.

Additionally, deliberation and reasoning have been effectively examined using the CRT. Such methods stem from the assertion that deliberative processing is more time consuming and cognitively demanding than intuitive processing (e.g., Kahneman, 2011). Accordingly, intuitive and deliberative thinking have been distinguished using time limits and/or cognitive loads. Two-response paradigms, for example, give participants two opportunities to answer a problem: at ‘Response 1’ while instructed to give the first response that comes to mind (sometimes under a time limit/secondary load constraint) and again at ‘Response 2’ without restrictions (Thompson et al., 2011, 2013). Participants who can reach the correct solution at Response 2 but not Response 1 are thought to have employed deliberative thinking.

Bago and De Neys (2019) combined measures of self-reported confidence and deliberation in a two-response paradigm with the CRT to examine the nature of correct responding (see also Thompson et al., 2013). In an exploratory analysis, they found that self-reported confidence was lower on deliberative trials (i.e., incorrect to correct answer change). However, they observed deliberation for only 8.93% of trials and, without any manipulation of confidence, could not rule-out issues of endogeneity (see Box 1). A low rate of deliberation is limiting because it is associated with low within-subject variance in thinking type which, in turn, limits subsequent analyses to largely between-subject comparisons that are susceptible to confounds. However, debiasing training paradigms can be used with the CRT to increase instances of deliberation (Isler et al., 2020) and elicit within-subject shifts in thinking type (Purcell et al., 2021). The current study builds on Bago and De Neys’ exploratory work by introducing an eye-tracked confidence index and a debiasing training paradigm.

Box 1: Endogeneity occurs when a predictor (X) may be affected by an outcome (Y) or both may be influenced by a third, usually unknown or unmeasured, variable (see Antonakis et al., 2010; Rohrer, 2018). When considering the relationship between confidence (X) and deliberation (Y) we must think on the possibility that deliberation will increase accuracy which may, in turn, influence confidence. In other words, endogeneity might be occurring because deliberation (Y) may be influencing confidence (X) via accuracy. Manipulating confidence independently of accuracy would reduce the likelihood of endogeneity, however, the manipulation must address task-specific confidence which is difficult to achieve without also impacting task-specific accuracy. Here our approach is to lean on a deception-free, within-participant, training paradigm in which we provide feedback and guidance. As training increases, we expect significant variance in confidence, for some, it may drop at first due to being notified of incorrect responding and then increase slowly as understanding increases. Importantly, this is expected to vary in a way that is similar but distinct from patterns of accuracy, which reduces the likelihood that our findings are dependent on an endogenous relationship between confidence and deliberation (via accuracy).

In addition to increasing within-subject variation in deliberation, debiasing training also reduces issues of endogeneity that have compromised previous examinations of confidence-deliberation relationship (see Antonakis et al., 2010). In a series of pre-registered studies, Isler et al. (2020) demonstrated that solution-specific methods like decision justification and feedback, but not time delay or memory recall, were effective methods for increasing instances of deliberation on the CRT (see also Purcell et al., 2021). Isler et al. postulated that manipulations were more effective when they involved specific reflection guidance. To illustrate, in contrast to time delay or recall, training involves providing task-specific feedback which can either decrease confidence by encouraging incorrect respondents to question their incorrect solution or increase confidence by validating correct solutions. Training is a highly task-specific manipulation and therefore a leading design option for manipulating confidence and subsequently reducing endogeneity when examining the confidence-deliberation relationship.

The current study builds on previous work testing the relationship between confidence and deliberation by examining whether confidence – tracked from the outset of reasoning –precedes deliberation, and whether this relationship is maintained in instances when task-specific confidence is manipulated via a training paradigm. We expected that self-reported and eye-tracked confidence would be associated with each other, that decreases in each index would be associated with an increase in the likelihood of deliberation, and that this relationship would be consistent across training. We also explored whether self-reported and eye-tracked confidence would account for unique variation in deliberation when considered simultaneously. Data and materials are publicly available[[2]](#footnote-3).

**Method**

**Participants**

Thirty-nine undergraduate psychology students at Macquarie University (Sydney, Australia) were awarded course credit for participation[[3]](#footnote-4). All participants had normal vision. Participants were 24 females and 14 males with ages ranging from 18 to 41 (*M*=21.24, *SD*=4.65)[[4]](#footnote-5).

**Apparatus**

The experiment was run on Experiment Builder 1.10.165 (SR-Research) and presented on a 24.5-inch LCD monitor (BenQ XL2540, refresh rate 240Hz, natural resolution 1920x1080) and right-eye movements were recorded with a desk mounted eye-tracker sampling at a rate of 1000 Hz (EyeLink 1000; SR Research Ltd., Osgoode, Ontario, Canada). Participants used a chinrest to maintain a viewing distance of 800mm.

A nine-point calibration was conducted to start and a one-point calibration prior to each item. If the one-point calibration was unsuccessful, the nine-point calibration was conducted again before the participant continued. The questions were presented with four multiple-choice options randomly allocated to corners of the screen (Figure 1D and 1F). Participants used a mouse to make their response. Areas of Interest (AOIs) were assigned to each multiple-choice alternative. AOIs were dynamic such that they reflected the response option not the screen location. Dwell, the amount of time participants looked at each AOI, was recorded for each trial during the Response 1 period (see Procedure).

**Materials**

Participants completed a training task shown to improve performance on CRT-like problems (Purcell et al., 2021, 2022). The task contained 3 no-lure practice items, 21 lure, and 21 no-lure items. Lure items were structured per the three original CRT questions (Frederick, 2005); however, the content and quantities were changed to prevent rote learning or recognition effects. Each lure item had a corresponding no-lure problem with a similar structure but no tempting incorrect response. Items reflected the word lengths of the original CRT item (+/-1 word). An example of a lure item based on the first CRT question is: “A shirt and a jacket cost $18 together in total. The shirt costs $10 more than the jacket. How much does the jacket cost?” (Answer: $4). The corresponding no-lure item was: “A phone and a wallet cost $1000 in total. The phone costs $800. How much does the wallet cost?” (Answer: $200).

**Procedure**

After providing consent, participants were given verbal and written instructions about the procedure and eye-tracking equipment. Two three-minute breaks were included throughout the task. After the practice items, participants were presented with seven blocks of six maths problems (three lure and three no-lure items), in the order: test block 1, training block 1, test block 2, training block 2, test block 3, training block 3, and test block 4. The order of the items was counterbalanced between blocks and randomised within blocks.

A two-response paradigm was imposed during test blocks (e.g., Thompson et al., 2011). Each problem was presented twice, once with a time limit (Response 1) and again with no time-limit (Response 2). A sample trial is displayed in Figure 1. Additionally, participants who did not respond before the Response 1 question timed out (Figure 1D) were reminded that the first time the question is presented they only have five seconds to respond and that they should respond with the first answer that comes to mind.



Figure 1*. Sample trial presentation. Screens A to C were presented for 3000ms, screen D for 5000ms or until a response was made, and E to G until a response was made. Some proportions are changed in this figure for readability.*

Training blocks differed from test blocks in that participants were only presented with the problem once and without a timing constraint. Participants were given feedback (correct or incorrect), presented with the full question and answer, and given a brief explanation as to why that was the correct solution (Purcell et al., 2021, 2022). For items and feedback see https://osf.io/ej3n2/.

**Results**

The following section includes a preliminary examination of the effect of the training manipulation on (1) performance and (2) confidence, and (3) the relationship between self-reported and eye-tracked confidence. The preliminary findings suggest that pattern of performance across training differs from the pattern of confidence across training, therefore, reducing concerns of endogeneity between confidence and deliberation via accuracy. The preliminary findings also show a positive association between the eye-tracked and self-reported measure of confidence indicating reasonable convergent validity. These findings illustrate a strong experimental foundation for examining our primary research questions. We then address our primary research questions about the relationship between (4) self-reported confidence and deliberation, (5) eye-tracked confidence and deliberation, and (6) the combined effects of self-reported and eye-tracked confidence on deliberation.

Thirty-nine participants completed three lure items in each of the four test blocks which were presented in two response conditions: Response 1 (timed) and Response 2 (untimed). Fourteen trials were not reported due to technical issues[[5]](#footnote-6), and 83 Response 1 trials timed out[[6]](#footnote-7), leaving 371 trials for analyses involving confidence. Self-reported confidence scores were ratings out of 100 provided immediately after Response 1. Eye-tracked confidence was calculated as the total dwell that occurred on the non-selected responses during Response 1 as a percentage of the maximum possible dwell time during Response 1 (5000ms); the less time a participant spent examining the alternative multiple-choice options, the higher their eye-tracked confidence[[7]](#footnote-8).

*Preliminary Analyses*

*(1)* *Training and performance.* To assess the effect of training on performance we conducted a repeated measures ANOVA with test block (T1, T2, T3, T4), problem type (lure, no lure), and response (1 or 2) included as predictors. Test block had a significant positive main effect on performance, (*p*<.001), indicating that training improved accuracy. Significant two-way interactions between problem type and test block *(p*<.001) and response and test block (*p*=.001), indicated that the effect of training was stronger for lure than no lure items, and for Response 1 than Response 2. No other effects were significant (see Appendix A1). Notably, improvements in performance were driven by improvements from T1 to T2 and T2 to T3 but not T3 to T4 (see Appendix, Table A1).

*(2)* *Training and confidence.* To assess the effect of training on confidence we conducted a repeated measures ANOVA with test block (T1, T2, T3, T4), problem type (lure, no lure) included as predictors. For self-reported confidence, problem type had a significant main effect – confidence was lower for lure than no lure items (*p*<.001) – and test block had a significant positive main effect (*p*<.022) – confidence increased over training. For eye-tracked confidence, only problem type was a significant predictor; confidence was lower for lure than no lure items (*p*<.001; see Appendix A2 for a full breakdown of these analyses). In contrast to performance, increases in self-reported confidence across test block were driven by differences between confidence at T1 and T4. Moreover, test block did not have a significant effect on eye-tracked confidence.

*(3) Self-reported and eye-tracked confidence.* A mixed model analysis with item nested within participant revealed that self-reported and eye-tracked confidence were significantly and positively associated; *F*(1,340.73)=13.29, *p*<.001.

Overall, our preliminary findings suggest that the training manipulation was effective in shifting both performance and confidence in a way that reduces concerns of endogeneity and increases the potential for within-subject comparisons of the confidence-deliberation relationship. The relationship between self-reported and eye-tracked confidence provides evidence of convergent validity between the measures.

*Primary Analyses*

Our primary analyses focus on the relationship between confidence and deliberation. A deliberation index was generated by pooling responses that indicated answer change (or not) from Response 1 to Response 2. First, items were coded to reflect change categories. Answers were coded to reflect participants’ accuracy (Correct or Incorrect) at Response 1 and Response 2 (see Table 1)[[8]](#footnote-9). Change codes were then used to form a binary deliberation index (yes, no). As in previous studies, answer change from incorrect at Response 1 to correct Response 2 was interpreted as an indication of the engagement of deliberative processes (Bago & De Neys, 2019; Thompson et al., 2011)[[9]](#footnote-10). As training increased, the likelihood of deliberation decreased (*p*<.001; for a full breakdown of the analysis of deliberation across test block, see Appendix A3).

Table 1. *Number of trials in each change category by test block.*

|  |  |  |
| --- | --- | --- |
| Test Block | Change Category |  |
|  | II | IC | CI | CC | Total |
| 1 | 42 | 30 | 4 | 15 | 91 |
| 2 | 18 | 25 | 0 | 43 | 86 |
| 3 | 12 | 16 | 4 | 62 | 94 |
| 4 | 25 | 12 | 1 | 62 | 100 |
| Total | 97 | 83 | 9 | 182 | 371 |

Note: *CC = correct answer at Response 1 and Response 2; II = incorrect answers at Response 1 and Response 2; CI = correct answer at Response 1 and incorrect answer at Response 2; IC = incorrect answer at Response 1 and correct answer at Response 2. Time-out trials have been excluded.*

Self-reported and eye-tracked confidence indices were lower for trials on which participants demonstrated deliberation (Figure 2). Binary logistic generalised linear mixed models with item nested within test block and test block nested within participant were used to analyse these findings in which the dependent variable was deliberation, and the predictors were item, test block and confidence.



 Figure 2. *Deliberation by (A) self-reported and (B) eye-tracked confidence. Confidence was lower for trials on which the participant engaged in deliberative thought. Error bars reflect +/-1SE.*

*(4) Self-reported confidence and deliberation.* For self-reported confidence, the model predicting deliberation revealed significant effects of item, *F*(2, 355)=11.27, *p*<.001, test block, *F*(3, 355)=5.02, *p=*.002, and self-reported confidence, *F*(1, 355)=19.77, *p*<.001. Follow up tests revealed that deliberation was more likely for items 1 and 3 than item 2 (see Appendix A4, Table A5), and that deliberation was less likely as test block increased (see Appendix A4, Table A6). The effect of self-reported confidence indicated that lower confidence was associated with a higher likelihood that the participant engaged in deliberative thinking (see Figure 2).

*(5) Eye-tracked confidence and deliberation.* For eye-tracked confidence, the model predicting deliberation revealed significant effects of item, *F*(2, 355)=8.68, *p*<.001, test block, *F*(3, 355)=5.32, *p=*.001, and eye-tracked confidence, *F*(1, 355)=6.11, *p*=.014. As above, deliberation was more likely for items 1 and 3 than item 2 (see A5, Table A7), and decreasingly likely as training increased (see Appendix A5, Table A8). The effect of eye-tracked confidence indicated that lower confidence was associated with a higher likelihood of deliberation (see Figure 2).

*(6) Self-reported and eye-tracked confidence on deliberation.* Finally, we included both self-reported and eye-tracked confidence indices in the one binary logistic generalised linear mixed model predicting deliberation. The model revealed significant effects of item, *F*(2, 354)=7.74, *p*=.001, and test block, *F*(3, 354)=4.59, *p=*.004, on deliberation. As above, deliberation was more likely for items 1 and 3 than item 2 (see Appendix A6, Table A9), and decreasingly likely as training increased (see Appendix A6, Table A10). The model indicated significant effects of self-reported confidence, *F*(1,354)=18.47, *p*<.001, and eye-tracked confidence, *F*(1, 354)=5.71, *p*=.017 on deliberation.

Overall, our primary findings suggest that as self-reported and eye-tracked confidence decreased, the likelihood of deliberation increased. They also suggest that the two indices accounted for unique variance in deliberation.

**Discussion**

Dual process models have proposed that metacognitions like confidence and uncertainty are involved in the engagement of deliberative thinking (e.g., De Neys, 2012; Thompson & Morsanyi, 2012). We tested this assertion using the CRT with a rigorous two-response training paradigm and measures of self-reported and eye-tracked confidence. The training component of the study yielded shifts in confidence during the study that differed from the shifts in performance. This difference lowers concerns of endogeneity. As expected, the self-reported and eye-tracked confidence measures were related to each other – demonstrating convergent validity. Importantly, we observed significant negative relationships between both measures of confidence and deliberation. Additionally, when examined concurrently the confidence indices both remained independent predictors of deliberation. These findings provide evidence for the core assertion that metacognitions, in particular low confidence, precede deliberation.

In line with Purcell et al. (2022), self-reported and eye-tracked confidence were related. However, the two measures made independent contributions to the engagement of deliberation. This independence could be due to the measures indexing different psychological phenomena. For example, in line with the three-stage model, eye movements and self-reported confidence may reflect distinct stages of processing related to the stage of conflict detection and the stage of analytic thought, respectively (Pennycook et al., 2015). However, it may also be due the nature of the measures. For example, eye-tracked confidence – based on the time participants spent looking at the non-selected multiple-choice alternatives – may have been related to self-reported confidence if Response 1 did not time-out at 5000ms. Future studies are needed to explore whether self-reported and eye-tracked confidence are indirect measures of the same phenomena or measures of separate metacognitive phenomena.

Consistent with Bago and De Neys' (2019) exploratory work, we found that lower self-reported confidence predicted a higher likelihood of deliberation. Following from the addition of a training component, the current study adds credibility to the previous finding by demonstrating that the relationship between confidence and deliberation was maintained when observing a higher proportion of deliberative trials (22.4%), greater within-subject variance in thinking type (97% of participants showed patterns of both intuitive and deliberative thinking), and reduced concerns about endogeneity. Moreover, because training is a form of task-specific metacognitive manipulation, the current findings provide the strongest support yet for the assertion that confidence may play a causal role in the engagement of deliberation.

This support is reinforced by the findings related to the time-course of the confidence-deliberation relationship; that is, whether metacognitive changes precede deliberation. Traditional self-reported confidence indices are limited because they are measured after ‘Response 1’. That is, after the cognitive constraint is removed and the reasoner can deliberate. This impacted the extent to which previous studies could assess the time course of the relationship between confidence and deliberation. In the current study the eye-tracked confidence measure was recorded before the time-constraint was removed. We note that despite our instructions to provide the “first response that comes to mind” and imposing a time-limit during Response 1 it is still possible that some participants may have engaged in deliberation prior to the deadline. However, it is with reasonable confidence that the negative relationship we observed between eye-tracked confidence and deliberation supports the assertion that lowered confidence precedes deliberation.

The negative relationship between confidence and deliberation can be explained by the contemporary dual process models. The logical intuition model and the three-stage model suggest that conflict triggers deliberation (De Neys, 2012, 2014; Pennycook et al., 2015). Under these models, eye-tracked and self-report confidence could be interpreted as inverse indices of conflict which precede deliberative thinking. The metacognitive model asserts that the fluency with which a response comes to mind impacts the reasoner’s affect, which cues deliberative thinking (Thompson et al., 2011, 2013). Accordingly, our eye-tracked and self-reported confidence indices may reflect fluency and metacognitive affect, respectively. Future projects may help to tease apart the components of metacognition (e.g., fluency and affect) as they pertain to deliberation by considering other indices of metacognition such as response times, pupil dilation or brain activity (e.g., via electroencephalography). On this note, we stress that – while the current article provides crucial support for the general assertion that metacognition is involved in the engagement of deliberation – future studies will be important for clarifying, operationalising, and testing potentially separate metacognitive mechanisms and their contributions to the engagement of deliberation.

Additionally, whether deliberation in the current study reflects the same phenomenon as the deliberation observed in minimal feedback or non-training studies warrants consideration (c.f. Boissin et al., 2021; Raoelison & De Neys, 2019). In dual process models, deliberative thinking is often conceived as encompassing many functions (Evans & Stanovich, 2013). It is, therefore, possible that the deliberative functions that occur as a result of spontaneous versus training-induced deliberation may differ in their dependence on, for example, inhibition, cognitive control, or hypothetical thought. On a similar note, we acknowledge that not all forms of deliberation will result in correct responses; our measure of deliberation requires that a person has changed from an incorrect response to a correct response and hence will not capture those cases where people deliberated but did not reach the correct response (e.g., some may have used deliberation to justify their initial response.) Future studies addressing the nature of deliberation may benefit from considering the roles of distinct cognitive functions underlying deliberation and the ramifications of specific deliberation prompts.

 Our findings support the contemporary dual process models of reasoning that emphasise the role of metacognition – specifically, the logical intuition model, the three-stage model, and the metacognitive model. As predicted, we demonstrated that eye-tracked and self-reported confidence were negatively associated with deliberation. The training component of the study strengthened this conclusion by yielding experimental shifts in task-specific confidence and thus lowered concerns of endogeneity. Moreover, the use of eye movements as an indicator of confidence provided novel support for the assertion that low confidence precedes deliberation and reduced concerns related to the use of self-reported confidence. These findings provide the strongest empirical support so far for the highly purported but under examined relationship between confidence and the engagement of deliberative thought.

REFERENCES

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *Leadership Quarterly*, *21*(6), 1086–1120. https://doi.org/10.1016/j.leaqua.2010.10.010

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

Boissin, E., Caparos, S., Raoelison, M., & De Neys, W. (2021). From bias to sound intuiting: Boosting correct intuitive reasoning. *Cognition*, *211*, 104645. https://doi.org/10.1016/J.COGNITION.2021.104645

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., 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., & 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

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

Isler, O., Yilmaz, O., & Dogruyol, B. (2020). Activating reflective thinking with decision justification and debiasing training. *Judgment and Decision Making*, *6*(15), 926–938.

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

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

Purcell, Z. A., Howarth, S., Wastell, C. A., Roberts, A. J., & Sweller, N. (2022). Eye tracking and the cognitive reflection test: Evidence for intuitive correct responding and uncertain heuristic responding. *Memory and Cognition*, *50*, 348–365. https://doi.org/10.3758/s13421-021-01224-8

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., & De Neys, W. (2019). Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem. In *Judgment and Decision Making* (Vol. 14, Issue 2). https://doi.org/10.17605/OSF.IO/6AEC3

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

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data: *Https://Doi.Org/10.1177/2515245917745629*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. Yale University Press.

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

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, *128*(2), 237–251. https://doi.org/10.1016/j.cognition.2012.09.012

**Appendix**

**A1. *Extended analysis of performance by test block, problem type and response.***

A three-way, repeated measures ANOVA was used to examine the effects of test block (T1, T2, T3, T4), problem type (lure, no-lure) and response (R1, R2) on performance. As stated in the main text of the article, averaging across the other factors, test block had a significant main effect on performance, *F*(1,99)=29.67, *p*<.001, *ηp2*=.473 (see Table A1 and Figure A1).

Table A1

*Simple effects of test block on performance. These results show the difference in performance between each test block. It demonstrates that performance was significantly greater from T1 to T2, and T2 to T3, but not T3 to T4 (see also, Figure A1).*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Block | Mean Difference | SE | *F* | *ηp2* | *p* |
| T2 – T1 | .48 | .08 | 11.27 | .255 | <.001 |
| T3 – T2 | .19 | .08 | 5.02 | .132 | .025 |
| T4 – T3 | -.06 | .06 | 8.17 | .198 | .317 |

*Note: Df*=33.

 Additionally, problem type had a significant main effect on performance, *F*(1,33)=124.35, *p*<.001, *ηp2*=.790; and response had a significant main effect on performance, *F*(1,33)=121.89, *p*<.001, *ηp2*=.787.



 Figure A1. *Performance by test block, problem type and response. Scores could range from 0-3. Error bars reflect +/- 1 SE.*

Additionally, there was a significant two-way interaction between problem type and test block, *F*(3,99)=10.80, *p*<.001, *ηp2*=.247 (see Table A2 for simple effects); there was also a significant two-way interaction between problem type and response on performance, *F*(1,33)=12.24, *p*=.001, *ηp2*=.271 (see Table A3 for simple effects). The three-way interaction between test block, problem type and response on performance was not significant, *F*(3,99)=1.23, *p*=.304, *ηp2*=.036. The results for performance across test block, problem type and response are presented in Figure A1.

Table A2

*Simple effects of problem type (No-Lure – Lure) by test block on performance. These results show the difference in performance between lure and no-lure problems at each test block. It demonstrates that performance was significantly greater for no-lure than lure trials at all test blocks but strongest at T1 (see also, Figure A1).*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test Block | Mean Difference*No-Lure – Lure* | SE | *F* | *ηp2* | *p* |
| T1 | 1.58 | .145 | 118.73 | 0.545 | <.001 |
| T2 | .93 | .134 | 48.17 | 0.327 | <.001 |
| T3 | .69 | .118 | 34.19 | 0.257 | <.001 |
| T4 | .90 | .152 | 35.06 | 0.262 | <.001 |

*Note: Df*=33.

Table A3

*Simple effects of response (Response 2 – Response 1) by problem type on performance. These results show the difference in performance for Response 2 (untimed) and Response 1 (timed) for both lure and no-lure responses. They demonstrate that performance was significantly greater for Response 2 than Response 1 for both lure and no-lure problems but that this difference was greatest for lure problems (see also, Figure 3).*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Problem Type | Mean Difference\**Response 2 – Response 1* | SE | *F* | *ηp2* | *p* |
| Lure | .78 | .088 | 78.56 | 0.704 | <.001 |
| No-Lure | .42 | .059 | 50.68 | 0.606 | <.001 |

*Note: Df*=33.

**A2. *Extended analysis of confidence by test block and problem type.***

Two-way, repeated measures ANOVAs were used to examine the effects of test block (T1, T2, T3, T4) and problem type (lure, no-lure) on confidence. For self-reported confidence, as stated in the main text of the article, problem type and test block were significant predictors; confidence increased over training, *F*(3,265)=3.27, *p*=.022, *ηp2*= 04 (see Table A4 for simple effects). Self-reported confidence was lower for lure (M=67.1, SE=2.62) than no lure items (M=80.6, SE=1.62); *F*(1,265)=70.16, *p*<.001, *ηp2*= 21. The interaction effect of problem type and test block on self-reported confidence was not significant, *F*(3,265)=0.013, *p*=.998, *ηp2*= 1.47e-04.

Table A4

*Simple effects of Test Block (T1, T2, T3, T4) on self-reported confidence. These results show the difference in self-reported confidence between each level of test block. They demonstrate that self-reported confidence was significantly higher at T4 than T1, but no other comparisons were significant.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Block | Mean Difference | SE | *t* | *P* |
| 1 – 2 | -1.37 | 2.29 | -.60 | .933 |
| 1 – 3 | -.90 | 2.29 | -.39 | .980 |
| 1 – 4 | -6.49 | 2.29 | -2.83 | .026 |
| 2 – 3  | .047 | 2.29 | 0.21 | .996 |
| 2 – 4  | -5.11 | 2.29 | -2.24 | .115 |
| 3 – 4  | -5.59 | 2.29 | -2.44 | .071 |

*Note: df=*748. Averaged across problem type.

For eye-tracked confidence, problem type was a significant predictor; confidence was lower for lure (M=4418, SE=20.2) than no lure items (M=4603, SE=20.2); *F*(1,265)=40.30, *p*<.001, *ηp2*= .13. Test block was not a significant predictor, *F*(3,265)=.83, *p*=.478, *ηp2*= 9.32e-03, nor was the interaction effect of test block and problem type, *F*(3,265)=.89, *p*=.445, *ηp2*= .01.

**A3. *Extended analysis for deliberation and test block.***

We conducted an exploratory analysis to assess the effect of training on deliberation for lure items. We fitted a logistic mixed model to predict deliberation (change codes: IC, CC|II) from test block (T1, T2, T3, T4), subject was included as a random intercept. The effect of test block was statistically significant and negative (beta = -0.39, 95% CI [-0.58, -0.20], *p* < .001). This suggests that as training increased, deliberation was less likely. Deliberation was significantly more likely at T1 than T3 (*p=*.028) and T4 (*p*<.001) but not T2 (*p*=.472). Deliberation was more likely at T2 than T4 (*p*=.001), but not significantly different from T3 (*p*=.136). Deliberation was not significantly different for T3 and T4 (*p*=.077). These results suggest that deliberation is more likely at T1 and T2 compared to T3 and T4. However, it should be noted that at T1 and T2 most non-deliberation trials are comprised of II change codes (incorrect at Response 1 and Response 2) while the majority of non-deliberation trials at T3 and T4 are comprised mainly of CC change codes (correct responses at Response 1 and Response 2). The distribution of change codes is shown in Figure A2.

 

Figure A2. *This figure shows how ‘change codes’ shift over test block – demonstrating the effect of training on performance and deliberation.* *Change codes: CC= correct at Response 1 and 2, IC=Incorrect at Response 2 and correct and Response 2, II = Incorrect at response 1 and 2, CI= Correct at Response 2 and incorrect and Response 2. XC and XI reflect time-out trials. Participants are ranked from highest (top) to lowest (bottom) by their total number of correct responses on lure items.*

**A4. *Follow up analyses for the model predicting deliberation from self-reported confidence, test block and item.***

Table A5*Follow-up analysis: The effect of item. Comparisons are coded according to the items (e.g., 1 reflects item 1 – those structured like the bat and ball problem in the original CRT). These findings show that deliberation was more likely for items 1 and 3 than item 2.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item  | *Coefficient* | *SE* | *t* | *p* |
| 1 - 2 | -1.820 | .38 | -4.77 | <.001 |
| 1 - 3 | -.533 | .29 | -1.79  | .074 |
| 2 - 3 | 1.308 | .41 | 3.16 | .002 |

*Df error = 354.*

Table A6

*Follow-up analysis: The effect of test block. Notably, the likelihood of deliberation was no different at T1 and T2, or at T3 and T4, however, deliberation was significantly more likely at T3 than T2.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Block  | *Coefficient* | *SE* | *t* | *p* |
| T1 - T2 | -.247 | .35 | -.714 | .476 |
| T1 - T3 | -1.014 | .37 | -2.71 | .007 |
| T1 - T4 | -1.296 | .39 | -3.30 | .001 |
| T2 - T3 | -.765 | .38 | -2.01 | .045 |
| T2 - T4 | -1.045 | .40 | -2.615 | .009 |
| T3 - T4 | -.280 | .42 | -.661 | .509 |

Note. *Df error = 354.*

**A5. *Follow up analyses for the model predicting deliberation from eye-tracked confidence, test block and item.***

Table A7

*Follow-up analysis: The effect of item. Comparisons are coded according to the items (e.g., 1 reflects item 1 – those structured like the bat and ball problem in the original CRT). These findings show that deliberation was more likely for item 2 than item 1, and item 3 than item 2.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item  | *Coefficient* | *SE* | *t* | *p* |
| 1 - 2 | -1.720 | .41 | -4.20 | <.001 |
| 1 - 3 | -.309 | .31 | -.99 | .324 |
| 2 - 3 | 1.415 | .43 | 3.29 | .001 |

Note. *Df error = 354.*

Table A8

*Follow-up analysis: The effect of test block. These findings show that deliberation was no more or less likely for test blocks T1 and T2, or T2 and T3, or T3 and T4. However, deliberation was significantly more likely at T1 than T3 or T4. Deliberation was also more likely at T2 than T4.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Block  | *Coefficient* | *SE* | *t* | *p* |
| T1 – T2  | .303 | .35 | .865 | .388 |
| T1 – T3 | .987 | .37 | 2.65 | .008 |
| T1 – T4  | 1.396 | .39 | 3.55 | <.001 |
| T2 – T3 | .682 | .38 | 1.77 | .077 |
| T2 – T4 | 1.092 | .40 | 2.71 | .007 |
| T3 – T4 | .410 | .43 | .964 | .336 |

Note. *Df error = 354.*

**A6. *Follow up analyses for the model predicting deliberation from self-reported confidence, eye-tracked confidence, test block and item.***

Table A9

*Follow-up Analysis: The effect of item. Comparisons are coded according to the items (e.g., 1 reflects item 1 – those structured like the bat and ball problem in the original CRT). These findings show that deliberation was more likely for item 2 than item 1, and item 2 than item 3.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item  | *Coefficient* | *SE* | *t* | *p* |
| 1 - 2 | -1.592 | .40 | -3.97 | <.001 |
| 1 - 3 | -.359 | .31 | -1.14 | .254 |
| 2 - 3 | 1.253 | .43 | 2.94 | .003 |

Note. *Df error = 354.*

Table A10

*Follow-up analysis: The effect of test block. These findings show that deliberation was no more or less likely for test blocks T1 and T2, or T2 and T3, or T3 and T4. However, deliberation was significantly more likely at T1 than T3 or T4. Deliberation was also more likely at T2 than T4.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Block  | *Coefficient* | *SE* | *t* | *p* |
| T1 - T2 | .303 | .35 | .855 | .252 |
| T1 - T3 | 1.016 | .38 | 2.67 | .008 |
| T1 - T4 | 1.244 | .41 | 3.06 | .002 |
| T2 - T3 | .709 | .39 | 1.82 | .070 |
| T2 - T4 | .955 | .41 | 2.35 | .019 |
| T3 - T4 | .246 | .43 | .571 | .568 |

Note. *Df error = 354.*

1. We acknowledge that the terms *confidence* and *uncertainty* have taken on slightly different meanings in the reasoning literature. For this article, in line with previous empirical treatments of these concepts, we consider them as the inverse of one another. For the most part, we use the term *confidence* as it closely reflects our operationalisation. [↑](#footnote-ref-2)
2. For materials, see <https://osf.io/9qbck/?view_only=ed0ca2aeab4a40c3a35466d956234ee2>. For data, see <https://osf.io/jnb8v/?view_only=29442b88181f46a58f46847dd0590989>. [↑](#footnote-ref-3)
3. This sample size was selected to obtain a similar number of total trials as Bago and De Neys (2019). Notably, we expected a greater number of deliberative trials, and within-subject variance in accuracy and answer than Bago and De Neys (2019) such that we would be sufficiently powered for our analysis. [↑](#footnote-ref-4)
4. One participant did not provide age or gender information. [↑](#footnote-ref-5)
5. These trials were not recorded – possibly due to failures in hardware/software. [↑](#footnote-ref-6)
6. Confidence was not recorded for ‘timed-out’ Response 1 trials because it was not sensible to calculate self-reported (“How confident are you in this response?”) or eye-tracked (the sum of dwell on non-selected responses) confidence if no response was selected. [↑](#footnote-ref-7)
7. This measure was based on the previously used dwell-based measure in (Purcell et al., 2022) and in line with literature relating dwell-time to the depth with which that information is processed (e.g., Glöckner & Herbold, 2011; Rayner, 1998). Here, the measure is created as the amount of dwell on non-selected responses as a percentage of the maximum possible dwell-time (the 5000ms participants had to respond) to create a form of reverse-score so that positive scores reflect greater confidence and to create a scale that is more easily compared to the 100-point self-reported confidence indices. [↑](#footnote-ref-8)
8. See Appendix, Figure A2 for a description of answer change across test block and participant. [↑](#footnote-ref-9)
9. Nine trials on which participants changed from correct to incorrect were excluded for this analysis because this category was difficult to interpret and likely to reflect random responding (see CI in Table 1). [↑](#footnote-ref-10)