

**The Transition from Effortful to Intuitive Reasoning:
Experience, Conflict and Working Memory Engagement.**

Zoe A. Purcell

B. Psychology (Hons.)

M. Res

Department of Psychology

Faculty of Human Sciences

Macquarie University, Sydney, Australia

This thesis is presented for the degree of Doctor of Philosophy (PhD)

August, 2019

Table of Contents

ABSTRACT.....	IV
STATEMENT OF CANDIDATURE.....	V
ACKNOWLEDGMENTS	VI
THESIS BY PUBLICATION	1
GENERAL INTRODUCTION	1
OVERVIEW	3
DUAL PROCESS THEORIES OF REASONING	4
<i>Working memory.....</i>	5
<i>Working memory and experience.</i>	6
<i>The development of dual process models.</i>	8
<i>The default-intervention dual process model.....</i>	10
<i>The logical intuition dual process model.....</i>	12
<i>The logical intuition model and conflict.....</i>	13
<i>The logical intuition model and domain-specific experience.</i>	17
<i>The logical intuition model and working memory.....</i>	21
SUMMARY AND AIMS.....	23
METHODS AND MEASURES	25
STRUCTURE OF THE THESIS.....	28
REFERENCES	29
PAPER 1	37
ABSTRACT	38
STUDY 1.....	47
<i>Method</i>	48
<i>Results.....</i>	50
<i>Discussion.....</i>	53
STUDY 2.....	56
<i>Method</i>	59
<i>Results.....</i>	63
<i>Discussion.....</i>	66
GENERAL DISCUSSION	67
<i>References.....</i>	71
PAPER 2	79
ABSTRACT	80
STUDY 1.....	88
<i>Method</i>	89
<i>Results.....</i>	94
<i>Discussion.....</i>	98
STUDY 2.....	99
<i>Method</i>	100
<i>Results.....</i>	102
<i>Discussion.....</i>	106
GENERAL DISCUSSION	108
<i>References.....</i>	114
SUPPLEMENTARY MATERIAL	119
PAPER 3	121
ABSTRACT	122

METHOD	134
RESULTS	140
DISCUSSION	147
<i>References</i>	153
APPENDIX A.....	160
APPENDIX B.....	161
APPENDIX C.....	162
PAPER 4.....	167
ABSTRACT.....	168
METHOD	179
RESULTS	184
DISCUSSION	194
APPENDIX	204
GENERAL DISCUSSION.....	209
RESEARCH AIMS AND PRIMARY FINDINGS	210
IMPLICATIONS FOR THE LOGICAL INTUITION MODEL	215
ALTERNATIVE EXPLANATIONS.....	219
STRENGTHS, WEAKNESSES, AND FUTURE DIRECTIONS	221
<i>Methodological issues</i>	222
<i>Integrating individual differences</i>	226
<i>The future of conflict</i>	227
CONCLUSION.....	229
<i>References</i>	232
ETHICS APPROVAL.....	234
APPENDIX	240
QUESTIONS DEVELOPED FOR PAPER 1 (STUDY 2)	241
QUESTIONS DEVELOPED FOR PAPERS 3 AND 4	261
VISUAL DESCRIPTIONS OF HYPOTHESISED RELATIONSHIPS	286

Abstract

This thesis examines and develops the logical intuition dual process model of reasoning. The logical intuition model stipulates that effortful reasoning processes can become intuitive with increased practice and experience. However, this assumption has not previously been tested, nor integrated with the model's key elements: conflict and working memory engagement. In six studies within four papers, this thesis addresses this gap by examining conflict, a form of cognitive uncertainty, and working memory engagement as determined by experience. By examining the relationship between mathematical experience and CRT performance, we first demonstrate that, as domain-specific experience increases, working memory dependence decreases. This transition is then used to examine the relationship between conflict and working memory engagement.

Paper 1 examines the effect of a secondary task on CRT performance across different levels of mathematical experience, demonstrating that working memory dependence decreases as experience increases. In Paper 2, we examine the relationship between explicit conflict and performance, and between explicit conflict and working memory engagement; finding that explicit conflict becomes a stronger predictor of performance as experience increased and that explicit conflict predicts working memory engagement. In Paper 3, using confidence ratings and novel eye-tracking measures, we demonstrate that participants register explicit and implicit conflict when giving incorrect responses on the CRT. In Paper 4, we find that implicit and explicit conflict factors independently predict working memory engagement.

These papers comprise the first empirical assessment of a core assumption in dual process reasoning theories: that reasoning processes can transition from effortful to automatic. In examining this transition, a novel approach is used to explore key elements of the logical intuition model of reasoning. Overall, findings support the logical intuition model, but they also introduce additional qualifications.

Statement of Candidature

I, Zoe Purcell, certify that the work in this thesis entitled “The Transition from Effortful to Intuitive Reasoning: Experience, Conflict and Working Memory Engagement.” has not been previously submitted for a higher degree to any other university or institution other than Macquarie University.

I also certify that the thesis is an original piece of research and it has been written by me. Any sources of information used throughout this thesis are acknowledged, including any help or assistance that I have received in my work and preparation of this thesis.

The research presented in this thesis was approved by the Macquarie University Human Ethics Review Committee, reference numbers: **5201701116** and **5201500347**.

Signed:

Zoe A. Purcell

25th August, 2019

Acknowledgments

I would like to acknowledge the people who were involved in this project. Firstly, to my principal supervisor, Colin Wastell, for his invaluable support these past years. I am grateful for his contribution to this thesis and to my experience as a post-graduate student. He was always available for advice and provided guidance when needed but also allowed this thesis to be my own work and to reflect my own interests and goals.

I would also like to express my gratitude to my associate supervisor, Naomi Sweller, for her unwavering support and guidance. Her attention to detail and constant positivity have been immeasurably helpful. I would also like to acknowledge my mentor, Stephanie Howarth, for her valuable contributions to the work in this thesis, but also for her reassurance and friendship. I would also like to thank Magda Osman, Andy Clark, Robert Logie and Simon Handley for their thought-provoking questions and valuable insights.

I would also like to express my gratitude to my family and friends for their unconditional support and encouragement. Particularly to Sinead Ferris who patiently and skilfully edited this thesis and was an invaluable sounding board throughout the project. Finally, to Kelsy Weavil, Lauren Shapiro, Andy Roberts, Bianca Slocombe, Kelsie Boulton and Olivia Green, thank you for your support and for making this time enjoyable and collaborative.

Thesis by Publication

This thesis has been prepared in the Macquarie University ‘Thesis by Publication’ format. Papers 1 through 4 have been written and prepared as independent publications. As such, there is some overlap in the literature cited and some unavoidable repetition across chapters, although it has been minimised as much as possible. The formatting of the papers within this thesis generally conforms to the Publication Manual of the APA, 6th Edition, although tables and figures are inserted within the manuscripts, to assist with readability of the thesis.

General Introduction

Overview

This thesis examines two core assumptions that underpin the logical intuition model of reasoning (De Neys, 2012, 2014). First, that reasoning becomes easier with practice, and second, that conflict, a form of cognitive uncertainty, is associated with the engagement of effortful reasoning. Although the first idea rests on sound foundations established in other areas of psychology, it has received little empirical consideration in the context of reasoning. While the second assumption has been studied in the context of reasoning, it has yielded inconsistent findings, particularly when concerning the well-known cognitive reflection test (CRT; Frederick, 2005). These assumptions together form an important feature of the logical intuition model. Elucidating this aspect of the model renders it more useful for generating testable predictions about reasoning. These predictions, when tested, deepen our understanding of reasoning and assist in the development of more accurate and valid models of reasoning. The current thesis incorporates theoretical and empirical considerations related to practice. It introduces novel methodologies that help to explain inconsistencies in previous studies and generate new findings that can direct the development and further examination of the logical intuition model.

This introductory chapter outlines several theories of cognition and reasoning that are most relevant to the empirical chapters that follow. It then addresses key methodological considerations and provides an overview of the thesis' structure. In line with the requirements for a thesis by publication, a comprehensive literature review is contained in the general introduction and the empirical papers. The current chapter provides context for the themes that are common across the empirical papers.

Dual Process Theories of Reasoning

The *Encyclopaedia of Cognitive Science* (Nadel, 2003) describes reasoning as the cognitive processes in which information is combined to yield new information, and problem-solving as a form of reasoning in which the transformation of information is directed towards a specific goal. This section focuses on a branch of reasoning and problem-solving research that stems from the ‘heuristics and biases’ research lead by Kahneman and Tversky in the 1970s (Kahneman, 2011). The heuristics and biases work demonstrated that people commonly violate logical principles when a task cues an alternative, intuitive response (see Table 1 for examples of bias tasks). To explain the common instances of “biased” responding, several models of reasoning emerged, including dual process theories (for an example of an early dual process theory see Wason & Evans, 1975).

Dual process theories of reasoning assert that there are two Types of reasoning: Type 1 (also known as System 1 or heuristic processing) that is fast and does not require working memory, and Type 2 (also known as System 2 or analytic processing) that is slow and requires working memory (e.g., De Neys, 2017b; Evans & Stanovich, 2013; Kahneman, 2011). The general explanation for the prevalence of “bias” was that people are cognitive misers who try to minimise cognitive effort where possible (Kahneman, 2011; Toplak, West & Stanovich, 2014). Type 2 processing is effortful, hence, when an individual is faced with a reasoning problem for which there is a low-effort solution, people will tend to opt for the default Type 1 solution – even in cases where the low-effort solution conflicts with the problem’s logical principles (see Table 1). Recent empirical developments have challenged this explanation and led to the development of several new dual process models (De Neys, 2017b). However, within those models, distinguishing between Types of reasoning by the involvement (or not) of working memory remains largely consistent.

Working memory.

Working memory is a central feature of many overarching models of cognition and reasoning. Given the multitude of definitions and perspectives of working memory, it is necessary to define the construct as it is used throughout this thesis. For this thesis, working memory can be conceived as a domain-general subsystem of the mind with a limited capacity that enables one to temporarily sustain a set of mental representations for further manipulations and processing (Carruthers, 2013). The subsystem is domain general in that it can sustain and integrate representations from multiple sensory cortical regions through attentional processes. Working memory is postulated as a core mechanism for a multitude of important psychological phenomena such as complex learning (Kyllonen & Dennis, 1996; Kyllonen & Stephens, 1990), creative thinking (Lee & Theriault, 2013), reading comprehension (Daneman & Carpenter, 1980; Turner & Engle, 1989), and even emotion regulation (Kleider, Parrott, & King, 2010; Schmeichel, Volokhov, & Demaree, 2008). Although the current thesis is concerned primarily with how and why working memory is engaged, it is important to consider what constitutes working memory.

The most prominent model of working memory is Baddeley's multicomponent model (Baddeley, 2012). For over 40 years, Baddeley and colleagues have reformulated the traditionally unitary concept of short-term memory (Atkinson & Schiffrin, 1968). Baddeley and Hitch (1974) proposed a three-component model that distinguished the phonological loop and visuo-spatial sketchpad, and a cognitive control centre – the central executive. Later, Baddeley added the episodic buffer component (Baddeley, 2000). The episodic buffer was assumed to hold “chunks” of information that contained integrated aspects from the phonological loop, visuo-spatial sketchpad and central executive, and connect working memory with long-term memory (Baddeley, 2000). Stemming from Baddeley and Andrade's (2000) link between phenomenological vividness, a person's subjective experience of mental imagery, and the episodic buffer, Baddeley (2000) asserted

that retrieval from the episodic buffer was associated with the binding of mental representations from multi-sensory sources and conscious awareness. Baddeley (2000, 2012) likened the episodic buffer to the “global workspace” in Baars’ (1988) model of consciousness.

Baars’ (1988) proposed a “global workspace” which facilitates the binding of information from distant cognitive sources. The global workspace theory asserts that when a mental representation enters the global workspace, it is broadcast to a multitude of cognitive resources, information and processes. The process of global broadcast is thought to coincide with the phenomenological experience of conscious awareness. Moreover, the combination of disparate ideas in the global workspace is thought to facilitate creative thought, learning and critical thinking. The link between working memory and the global workspace was also adopted in Carruthers’ (2013, 2015) overarching model of cognition.

Carruthers (2015) suggested that the pattern of activation of a reasoning process determines whether that process is experienced consciously or non-consciously. That is, he drew a distinction between types of reasoning that are slow and conscious, and those that are fast and non-conscious and asserted that these processes are distinguished by the nature of processing rather than physical or mechanistic differences. The structures underlying fast and slow processes are not separate mechanisms, rather, ‘slow’ processes are the result of highly active ‘fast’ processes. The increased activation of non-conscious processes can lead to a process’ mental representation being held in working memory and globally broadcast. As such, Carruthers developed an architecture for reasoning that offers a cognitively tractable framework for the prominent dual process models of reasoning (e.g., Kahneman, 2011).

Working memory and experience.

In addition to providing a cognitive framework for dual process models, the literature on working memory can be used to consider the role of experience in influencing

the transition of processes from Type 2 to Type 1. Literature from the working memory field suggests that as domain-specific experience increases, the dependency on working memory to complete tasks within that domain decreases (e.g., Ericsson & Kintsch, 1995; Guida, Gobet, Tardieu, & Nicolas, 2012). The most prominent explanation for this phenomenon is Ericsson and Kintsch's (1995) long-term working memory model. This account suggests that skilled individuals can use long-term memory in a manner that allows them to circumvent some of the limitations of working memory (Chase & Ericsson, 1982).

The long-term working memory theory asserts that, due to greater domain-specific knowledge, highly skilled individuals have developed advanced encoding and retrieval structures that facilitate the use of long-term memory for temporary storage. Using long-term memory for an ongoing task has two advantages. First, information stored in long term memory is protected from the detrimental effects of secondary or concurrent tasks. Second, using long-term memory stores alleviates the need for working memory resources such that they can be allocated to secondary tasks, or used for more complex binding or decoupling processes. These ideas are explored in more depth in the empirical sections of this thesis. However, it is important to note at this point that it is widely accepted in the working memory literature that, as an individual's domain-specific experience increases, the associated processes can become automated. That is, the dependency of those processes on working memory decreases to such an extent that the task can be completed successfully with limited cognitive resources. In other words, a Type 2 process can decrease its dependence on working memory, effectively becoming a Type 1 process. This trajectory of reasoning processes from Type 2 to Type 1 can be used to examine dual process models of reasoning.

The development of dual process models.

Dual process models, particularly the default-intervention model, have their roots in Evans' early work on the heuristic-analytic model (Evans, 1984). The heuristic-analytic model contained several ideas that are still prominent in modern dual process theorising. Evans (1984, 1989, 2006) originally proposed the model to explain the prevalence of biased responding on popular reasoning paradigms (see Evans, 1989). Like the examples in Table 1 (Paradigms 1-3), these paradigms often presented problems in which the logical or normatively correct answer conflicted with beliefs and prior knowledge. Evans' (1984, 1989) original heuristic-analytic account suggested that biased responding may result from the omission of relevant information or the inclusion of irrelevant information in the mental representation of the problem, at the heuristic stage of reasoning. Analytic processing was thought to act on the representations of the problem generated at the heuristic stage, hence, if the heuristic representation was biased, the analytic processes would subsequently be biased.

Table 1. *Common paradigms used to explore dual reasoning.*

1. *Conjunction fallacy task:*

Sarah is 12 years of age. She is very talkative and sociable. She goes to drama classes and is learning to play the guitar. She wants to be a pop singer or an actress. Which one of the following statements is more likely?

1. Sarah likes to cook*
2. Sarah likes to cook and she collects pop magazines

2. *Belief bias task:*

All living things need water

Roses need water

Therefore, roses are living things

1. The conclusion is valid

2. The conclusion is invalid*

3. *Base-rate neglect task:*

In a study 1000 people were tested. Among the participants there were 995 nurses and 5 doctors. Pat is a randomly chosen participant of this study.

Pat is 34 years old and lives in a beautiful home in a posh suburb.

Pat is well spoken and very interested in politics and invests a lot of time in his or her career. Which of the following is more likely?

1. Pat is a nurse*

2. Pat is a doctor

4. *The bat and ball problem:*

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

1. 5 cents*

2. 10 cents

Note: The normatively correct response is indicated with an asterisk.

Evans introduced the idea that analytic processes could play a supervisory role and intervene upon inaccurate heuristic processing (Evans, 2006). Evans asserted that the heuristic system retrieves a default mental representation that generates default responses, inferences, or decisions. The analytic system may or may not intervene to modify or change the mental representation. Intervention, Evans (2006) suggested, is dependent on contextual factors like time constraints as well as individual factors like cognitive ability or thinking style. Thinking style reflects whether a person is predisposed by personality or cultural influence to engage in analytic thinking. These ideas are maintained in the contemporary default-intervention model (Evans & Stanovich, 2013). Although the heuristic-analytic model was heavily criticised for issues such as unfalsifiability and vague

definitions (e.g., Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011; Osman, 2004; Keren, 2013), it marked a pivotal point in the development of reasoning theories.

Since Evans' (2006) reformulation of the heuristic-analytic model, several theories of reasoning have emerged that focus on the distinction between Types by their cognitive features, such as working memory involvement, rather than the problem's features, such as incongruence between the believability and logicity of a problem (e.g., see Table 1, Paradigm 2). These contemporary theories include the default-intervention model (Evans & Stanovich, 2013) and the parallel-competitive model (Epstein, 1994; Sloman, 2014; Sloman, 1996). Recently, hybrid models have been developed that combine elements of the default-intervention and parallel-competitive theories. These hybrid models include the parallel processing model (Handley & Trippas, 2015), the logical-intuition model (De Neys, 2012, 2014), the three-stage model (Pennycook, Fugelsang, & Koehler, 2015) and the metacognitive model (Thompson, Turner, & Pennycook, 2011). Although the studies in this thesis could be applied to many of the dual process theories, they focus on the implications for the default-intervention model due to its prominence across several fields of psychology, and the logical intuition model due to its capacity to generate predictions about the interaction between practice and conflict, explained in more detail below. Therefore, the following sections focus on these two models of reasoning.

The default-intervention dual process model.

The default-intervention dual process model presented by Evans and Stanovich (2013) maintains the assertion that the two Types operated serially. That is, when faced with a problem, individuals typically instigate an initial (default) Type 1 process, but that this default process can be subsequently overridden by Type 2 processes (intervention) if required. The default-intervention perspective clarifies the previously vague, and therefore difficult to operationalise, definitions of Type 1 and Type 2 processes. Evans and Stanovich assert that the defining characteristics of Type 1 processes are autonomy and

that they do not require working memory, and for Type 2, that they involve cognitive decoupling and do require working memory.

The updated default-intervention account also emphasises the fast and effortless nature of solutions that are based on beliefs and prior knowledge, and the idea that these solutions would be brought to mind before a logical output could be produced. This assertion is supported by studies that demonstrated detrimental effects of memory loads and time constraints on logical responding (e.g., Evans & Curtis-Holmes, 2005; Johnson, Tubau, & De Neys, 2014). For example, in a study using the base-rate neglect task, Evans and Curtis-Holmes (2005) demonstrated that participants were more likely to produce belief-based responses when placed under a time constraint. However, this study had only 50 participants and a single bias task with a forced-choice response format. Moreover, while logic-based responses were reduced under time-constraint, they were not entirely eliminated indicating that some participants were able to reach the logical solutions under constraint. Recent studies with larger sample sizes and multiple bias tasks found that many participants were able to reach logical solutions in their first, immediate response while under time and memory load constraints. For example, Bago and De Neys (2017) frequently observed correct immediate responses in a study that employed base-rate and syllogistic tasks across four experiments with sample sizes of 99 to 115 participants. These recent findings suggest individual differences may play an important role and suggest that previous studies, such as the Evans and Curtis-Holmes (2005) study, may not be as supportive of the the time course assumption in the default-intervention model as originally thought.

The default-intervention model asserts that Type 2 processes, responsible for intervention, demand working memory resources (Evans & Stanovich, 2013). However, several studies have demonstrated that participants provide incorrect problem solutions to ‘bias tasks’ despite the involvement of working memory. For example, when participants

are provided the opportunity to rethink their initial incorrect response, they do not typically change their response (e.g., Shynkaruk & Thompson, 2006; Thompson, Prowse Turner, & Pennycook, 2011). However, they do demonstrate changes in confidence, which indicates that additional processing, presumably involving working memory, was executed (see Thompson et al., 2013). To account for studies indicating that working memory engagement does not ensure intervention, Evans and Stanovich (2013) suggest that the involvement of working memory does not necessitate correct responding. Moreover, they argue that intervention may be dependent upon additional factors such as the problem's novelty and difficulty, and the person's motivation. However, the model does not explain the cognitive mechanisms that underpin the relationship between these factors and engagement of Type 2 processes or possible intervention. Several dual process models have been developed that make more precise assertions about the determinants of Type 2 processing than the default-intervention model. The logical intuition model, for example, presents a comprehensive explanation for the interaction between Type 1 and 2 processing and the determinants for the engagement of Type 2 working memory dependent processes.

The logical intuition dual process model.

The logical intuition model developed by De Neys (2012, 2014; Bago & De Neys, 2017a) posits that logical principles can be enacted in Type 1 processing and that conflict detection is a key determinant for the engagement of Type 2 processes. De Neys argues that Type 1 processes can generate both stereotypical or belief-based responses, and logical or probabilistic responses. Multiple Type 1 responses may be generated in parallel, but not necessarily with equal strength or salience. While the process with the strongest salience is likely to “win out”, alternative (potentially logical) processes can affect the reasoning process. That is, conflict between alternate Type 1 processes may occur if more than one response is activated by a task. If conflict is detected, Type 2 processes may be engaged

and, in some cases, an alternative solution may be selected over the process that possessed the highest salience at the outset of processing.

In bias tasks, incorrect responses are commonly observed (see Table 1 for examples). De Neys (2012) suggests that incorrect responding on bias tasks may indicate that the processes leading to stereotypical responses had the strongest initial activation. However, he postulates that logical processes may also be engaged at the outset of reasoning. In other words, logical and probabilistic knowledge may still influence reasoning, even if these solutions are not ultimately enacted. The simultaneous activation of multiple Type 1 processes, regardless of their basis in belief or logic, is thought to generate cognitive conflict.

The logical intuition model and conflict.

The suggestion that competition between Type 1 processes can generate conflict is supported by research that has measured participants' conflict detection on bias tasks using several indirect measures such as confidence and response latencies (see De Neys, 2012 for a review). Conflict detection studies typically compare lure (also known as classic, standard or conflict versions) and no lure items (also known as control or no conflict versions). Lure items, like the bias tasks presented in Table 1, cue two common responses. No lure items have been developed to mirror several bias tasks but cue only a single response (e.g., Bonner & Newell, 2010; De Neys & Glumicic, 2008). An example of a lure and no lure version of the base-rate neglect task are presented in Table 2. By comparing lure and no lure items, conflict detection studies have been able to examine the effects of problems with multiple cued responses versus problems with a single cued response. That is, they compare problems that may induce competition between Type 1 processes, and problems that are not likely to generate multiple responses or consequent competition. If logical principles are not taken into account by the heuristic respondents, then the contrast

between heuristic responding on lure items and correct responding on no lure items should not have an impact on reasoning.

Table 2. *Example of a lure and no lure version of the base-rate neglect paradigm.*

<i>Lure version:</i>	<i>No lure version:</i>
A psychologist wrote thumbnail descriptions of a sample of 1000 participants consisting of 995 females and 5 males . The description below was chosen at random from the 1000 available descriptions.	A psychologist wrote thumbnail descriptions of a sample of 1000 participants consisting of 995 males and 5 females . The description below was chosen at random from the 1000 available descriptions.
Jo is 23 years old and is finishing a degree in engineering. On Friday nights, Jo likes to go out cruising with friends while listening to loud music and drinking beer.	Jo is 23 years old and is finishing a degree in engineering. On Friday nights, Jo likes to go out cruising with friends while listening to loud music and drinking beer.
Which of the following two statements is more likely?	Which of the following two statements is more likely?
a. Jo is a man ⁺	a. Jo is a man ^{*+}
b. Jo is a woman [*]	b. Jo is a woman

Note. *The lure version cues a heuristic response that conflicts with the logical principles. * = logical response according to probability theory principles, + = heuristic response. The lure version contains multiple response cues, whereas the no lure version cues only a single response.*

Conflict detection studies have demonstrated differences between lure and no lure items on a number of measures thought to reflect cognitive conflict. These results have been interpreted as an indication that ‘biased’ reasoners (i.e. those giving the heuristic answer to lure problems) register that their response conflicts with logical principles (e.g., De Neys, 2008). The indicators of conflict can be considered on a continuum of awareness, from implicit (low awareness) to explicit (high awareness). Evidence of conflict has been

observed via implicit indicators, for example, participants show higher autonomic arousal (De Neys, Moyens, & Ansteenwegen, 2010), higher activation in the anterior cingulate cortex (De Neys, Vartanian, & Goel, 2008; Simon, Lubin, Houdé, & De Neys, 2015), and longer response times (e.g., Frey, Johnson, & De Neys, 2018; Thompson, Striemer, Reikoff, Gunter, & Campbell, 2003).

Conflict has also been measured using explicit indicators such as confidence in one's response (e.g., De Neys, Cromheeke, & Osman, 2011; Hoover & Healy, 2019) and feelings of error or 'rightness' (e.g., Gangemi, Bourgeois-Gironde, & Mancini, 2015; Thompson et al., 2011). Despite these self-reported indications of conflict, biased participants rarely mention the task's logical components in think-aloud protocol designs (see De Neys & Glumicic, 2008). This absence of vocalising suggests that although there are explicit self-reported symptoms of conflict, there is a limit to the level of awareness that individuals have of the conflicting logical principles. Conflict detection effects have also been observed under conditions where Type 2 processes have been experimentally suspended, for example via secondary tasks and time constraints (De Neys, 2017a). Therefore, conflict is thought to affect Type 1 processing. Generally, conflict detection studies support the assertion that logical principles affect Type 1 reasoning processes even when individuals provide heuristic responses. However, not all studies examining conflict on bias tasks have found evidence for conflict detection (e.g., Travers, Rolison, & Feeney, 2016).

The CRT is a three-item bias task that contains the well-known bat and ball problem (see Table 1, Paradigm 4). Several studies have examined whether individuals solving the bat and ball problem experience conflict (e.g., Bago, Raoelison, & De Neys, 2019; Mata, Ferreira, Voss, & Kollei, 2017). However, they have yielded inconsistent findings and, subsequently, drawn different conclusions. For example, Travers et al.'s (2016) mouse-tracking study did not observe conflict effects, nor did Mata et al.'s (2017)

eye-tracking study. Travers et al. (2016) found evidence to suggest that correct responding on the bat and ball problem was associated with a pattern of mouse-movements to suggest a default-intervention strategy. That is, respondents who gave correct responses were more likely to move their cursors towards the heuristic (10 cents) option before moving to select the correct (5 cents) option. This was interpreted as an indication that the participants had considered the heuristic option, potentially via Type 1 processes, before overriding that solution process and selecting the correct option. When comparing heuristic responding on lure items to correct responding on no lure items, Travers et al. (2016) did not find evidence for longer response times, or differences in cursor trajectories. That is, they did not find evidence for conflict detection. Travers et al. (2016) concluded that their study supported the default-intervention interpretation of bias on the bat and ball problem.

Mata et al. (2017) examined participants' eye movements when solving the bat and ball problem. They found that correct respondents attended to the lure version of the problem to a greater extent than the no lure version, whereas heuristic respondents did not attend to the lure version to a greater extent than the no lure version. This was interpreted as an indication that errors on the bat and ball problem may stem from inaccurate interpretation and representation of the problem at an early stage of processing. Hence, they asserted that this finding supported a two-stage model of reasoning as postulated in Evans' heuristic-analytic model (Evans, 1984, 1989, 2006; see also Mata, Schubert, & Ferreira, 2014). Additionally, though not the primary aim of the paper, they did not find evidence for conflict effects when comparing the heuristic responding on lure items to the correct responding on no lure items. However, as the authors noted, the study may have been underpowered to detect small effects regarding the conflict sensitivity analysis.

This concern sparked a discussion about Mata et al.'s (2017) results in a commentary by Frey, Bago, and De Neys (2017) and a response by Mata and Ferreira (2018). Frey et al. (2017) noted that the lack of conflict sensitivity was based on a null

finding from frequentist statistics and, therefore reanalysed the data using a Bayesian analysis. The Bayesian approach found weak support for the null hypothesis, leading Frey et al. (2017) to classify the finding as “merely anecdotal” (p. 2) and suggest that more data was needed before a conclusion could be made in regards to conflict sensitivity on the bat and ball problem.

Mata and Ferreira’s (2018) response to the commentary accepted this qualification but added that further work was needed to assess when and why heuristic respondents show conflict sensitivity. Given that conflict sensitivity should only occur when the relevant biased and logical principles are intuitive, they postulated that experts may not exhibit signs of conflict if they do not share the biased intuitions of laymen (Obersteiner, Hoof, Verschaffel, & Dooren, 2016; Svedholm-Häkkinen, 2015), and that children may not show conflict sensitivity if they do not possess the knowledge-based biased intuitions (De Neys & Vanderputte, 2011; Jacobs & Potenza, 1991). The suggestion that expertise may moderate conflict sensitivity could also be applied to the lack of conflict sensitivity observed in Travers et al.’s (2016) study. The moderation by expertise can also be conceived in a more nuanced approach by incorporating the logical intuition model’s concept of relative activations (Bago & De Neys, 2017a). That is, rather than the expert or inexperienced reasoners not possessing one of the competing intuitions, they may exhibit less conflict sensitivity because the competing intuitions have unequal levels of activation.

The logical intuition model and domain-specific experience.

The logical intuition model proposed a relationship between the strength or salience of Type 1 processes and conflict (Bago & De Neys, 2017a; De Neys, 2012, 2014; for similar views see Pennycook et al., 2015; Trippas & Handley, 2017). As mentioned above, the logical intuition model suggests that multiple Type 1 responses may be initiated in response to a task and that the relative strength of the processes determines the generation of conflict (Bago & De Neys, 2017a). This proposition forms the basis for the predictions

made in the current thesis regarding the relationship between domain-specific experience and conflict.

Figure 1 presents the hypothesised relationship between competing processes and conflict. The four graphs reflect the activations of a heuristic process and a logical process at a time point along a continuum of increasing experience in the domain of the logical principle. As domain-specific experience increases, the activation of the logical intuition increases, eventually surpassing that of the heuristic process. The process with the highest activation will typically win out; at points 1 and 2, the heuristic response is most likely to be given, and the logical response at points 3 and 4. That performance changes with training is not very interesting. However, the implications of this shift for conflict detection are intriguing.

As the logical process is trained, its activation increases. At the initial time point, the heuristic process is much more salient than the logical process. This should result in low levels of cognitive conflict and less chance of conflict detection. At the intermediate time points (2 and 3) the relative activations of the heuristic and logical processes are similar and as such should generate greater conflict and subsequent conflict detection. Finally, after sufficient increases in domain-specific experience, the activation of the logical process should become greater than the heuristic process and their relative activations more unequal. Therefore, less conflict should be generated and detected. This proposed interaction between domain-specific experience, relative activations, and conflict is examined more closely in the empirical chapters of this thesis.

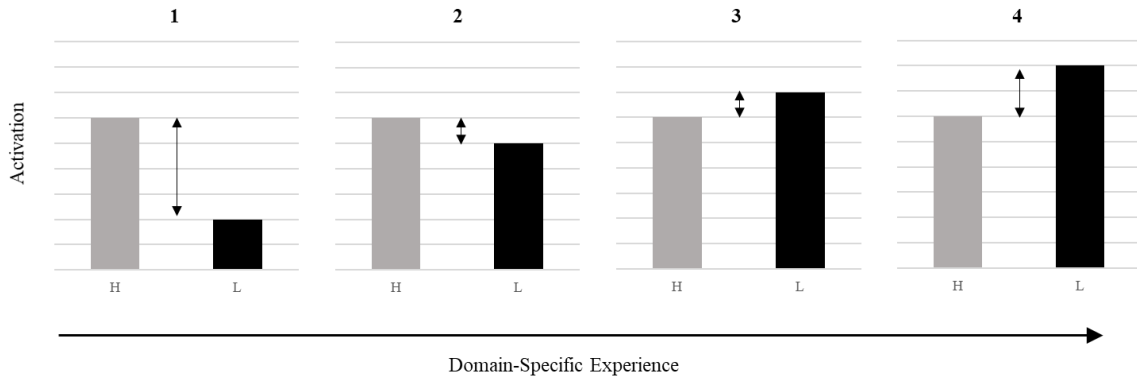


Figure 1. *Conjectural model of the relationship between domain-specific experience and the relative activations of competing processes. H=Heuristic process, L=Logical process.*

Studies have attempted to manipulate the salience of logical intuitions by changing the task properties (Bago & De Neys, 2017b, 2019). Bago and De Neys (2019) manipulated the extremity of base-rates to present a stronger or weaker cue for the logical intuition (see also Pennycook, Fugelsang, & Koehler, 2012; Pennycook et al., 2015). For example, the base-rate task in Table 2 might be altered to say 997 women and 3 men for an extreme base-rate, or 700 women and 300 men for a moderate base-rate. Bago and De Neys (2019) demonstrated that reducing the extremity of the base-rate decreased the number of logical (base-rate) responses. Additionally, participants who gave immediate heuristic responses were less likely to register the conflict (in this case, indicated by changes in confidence), whereas participants who gave immediate logical responses were more likely to register conflict. This study presented an elegant test of the interaction between relative activations and conflict proposed in the logical intuition model but, as the authors note, the hybrid model is a “work in progress” (p. 23) with many significant empirical and theoretical challenges ahead.

Two key issues raised by Bago and De Neys (2019) include a general lack of a priori predictions for salience manipulations and, relatedly, a lack of specification for what constitutes the “strength” of a process. In an earlier study by Bago and De Neys (2017b)

they hypothesised that the order in which information was presented would influence the relative strength of the intuitions. Based on Pennycook et al. (2015), they predicted that the information that was presented most recently would be most salient and therefore increase the strength of the corresponding intuition. However, when Bago and De Neys (2017b) incorporated the order manipulation with a two-response paradigm, they found the opposite effect of order manipulation for the first of the two responses provided by the participants. The expected pattern did appear for participants' second responses.

Accordingly, Bago and De Neys suggested that the intuition corresponding to the most recent piece of information presented may not have reached its peak activation until the participants were giving their second response. Bago and De Neys note (2019) that while this interpretation may assist in developing a deeper understanding of the rise and fall of intuition activations, it was not conducted on the basis of an a priori hypothesis.

Generating a priori hypotheses is a much simpler task when the constructs to be examined have been well defined. Bago and De Neys (2019) acknowledged this issue and note that they employed the term "strength" as a "functional label to refer to the hypothesised activation level of an intuitive response" (p. 24). This definition is somewhat circular and offers little by way of explaining what may underlie "strength". However, the acknowledgement did present an opportunity for other authors to qualify and examine what may constitute a process's strength.

In this thesis, "strength" or "salience" is thought to be derived from the amount of practice that a process has undergone. That is, the more experience an individual has in employing a solution process, the greater the "strength" of that process will be when it is cued by a task or situation. This is in line with De Neys's (2012) suggestion that a person's mastery of a normative principle is reflected in a heightened activation state when the person is presented with a problem. Thompson et al.'s (2011) metacognitive dual process approach considers a similar explanation for the strength of a process as it relates to

fluency. It asserts that implicit cues, such as the ease with which a memory is retrieved, affect the fluency of a response. Lower fluency, they suggest, is associated with a lower ‘feeling of rightness’. The ease with which a memory is retrieved, however, is thought to stem from the individual’s domain-specific experience, such that the more familiar the cue, the easier the process is brought to mind (Benjamin & Bjork, 2014). In reasoning and decision making, the cues within bias tasks are thought to cue different processes, and the processes that are cued are thought to have different levels of salience. This thesis highlights the potential for salience to be considered as determined by practice and expertise.

Several theorists have alluded to the idea that expertise may affect salience. The very notion that we have logical intuitions is based on the idea that processes can become automatised through repeated exposure and experience (De Neys, 2012, 2014). The model asserts that, through experience, logical processes can become intuitive such that they no longer require working memory. However, beyond its role in developing logical intuitions, domain-specific experience has received little theoretical consideration within the logical intuition model. This thesis provides a more detailed proposition for the role of domain-specific experience as it relates to the “strength” of processes and subsequent conflict generation and working memory engagement.

The logical intuition model and working memory.

A primary assertion in the logical intuition model is that the detection of conflict determines the engagement of working memory dependent, Type 2 processes (De Neys, 2012; for a similar suggestion see Pennycook et al., 2015). This assertion has been examined using cognitive constraint paradigms that temporarily suspend the individual’s capacity to engage working memory resources. These paradigms were developed on the basis that working memory has a limited capacity and operations requiring working memory are time-consuming. Accordingly, the most widely employed cognitive constraint

techniques involve secondary tasks that reduce the individual's available working memory capacity (e.g., Johnson, Tubau, & De Neys, 2016), and time constraints that reduce the individual's capacity to employ time-consuming processes (e.g., Thompson et al., 2013). These constraints have been employed in conjunction with conflict detection measures to test the relationship between conflict and working memory engagement (Bago & De Neys, 2017a).

In a comprehensive article containing two bias tasks (base-rate and syllogistic) and four experiments, Bago and De Neys (2017a) examined the relationship between conflict and working memory engagement. The study employed a two-response paradigm in which participants are given two chances to answer the problems (Thompson et al., 2011, 2013), in this case once with a time-limit and cognitive load, and again, with no restrictions. Bago and De Neys (2017a) demonstrated that lowered confidence ratings were associated with answer change. That is, participants who gave a different response on their first attempt, under constraints, to that which they gave on their second attempt, without constraints, showed lower confidence ratings at the initial response than those who did not change their response between attempts. Although they observed conflict detection by participants who did not change their response, the detection was more pronounced in instances when participants changed their answers. This was interpreted as an indication that conflict detection, as a function of relative strengths of competing activations, was associated with working memory engagement. This thesis builds on these findings, employing cognitive constraints and measures of conflict to examine the assertion that conflict is involved in the engagement of working memory dependent, Type 2 processing.

The key factors examined in this thesis are domain-specific experience, conflict and working memory engagement. Predictions for how these three factors relate to one another can be formed by combining elements from the working memory literature and the logical intuition model. Models of working memory suggest working memory is critical for

learning. The long-term working memory model suggests that as expertise increases, working memory dependence decreases. This suggests that when mastering a new reasoning process, an individual will experience greater working memory engagement during learning. However, after enough domain-specific experience is gained, their dependence on working memory to employ the learnt reasoning process will decrease. Thereby, the trajectory of increasing domain-specific experience can be expected to elicit an initial increase in working memory engagement followed by a decrease in working memory engagement.

When this trajectory of working memory engagement across increases in experience is considered in dual process terms, a parabolic function emerges. That is, as experience increases, reasoning should transition from Type 1 (employing an incorrect solution procedure, i.e. prior to learning) to Type 2 (i.e. during learning), and back to Type 1 (employing the correct procedure with ease, i.e. after automation). The logical intuition model suggests that conflict is associated with working memory engagement. Therefore, the same parabolic function should emerge for the trajectory of conflict across experience. That is, the reasoner should experience little conflict prior to learning (associated with Type 1 processing), greater conflict during learning (associated with Type 2 processing), and little conflict after automation (associated with Type 1 processing). This proposed trajectory of automation and its ramifications for working memory engagement and conflict form the basis of the research questions, predictions and experimental designs employed in the empirical papers presented in this thesis.

Summary and Aims

The logical intuition model presents a detailed model for the momentary engagement of Type 2 processes. However, in its current state, the model does not include a thorough account for the interaction between domain-specific experience and the key elements of the model: conflict and working memory engagement. The literature on

working memory and expertise can be readily incorporated into the logical intuition model, though, and their combination facilitates the generation of clear predictions. While the logical intuition model rests on the assumption that reasoning processes, even logical or probabilistic ones, can become automatised, this proposal has not been directly examined (De Neys & Pennycook, 2019).

The logical intuition model asserts that the relative “strengths” of intuitive processes determine the detection of conflict, however, as previous authors have noted, the determinants for “strength” are not yet well defined. The current thesis explores the possibility that the salience of an activated process may be determined by an individual’s experience in the relevant domain. Subsequently, the thesis examines the interaction between changes in process salience, elicited by increases in domain-specific experience, and consequent conflict and working memory engagement. The idea that practice and learning can change the nature of reasoning processes is not novel, however, the application of this idea to empirically test the assumptions of models of reasoning is a new approach.

The logical intuition model and its hybrid counterparts are in their infancy and arguably present more questions and hypotheses than answers or explanations (see De Neys, 2017b; De Neys & Bago, 2019). The current thesis aims to contribute to the growing number of studies testing the logical intuition model by investigating four primary questions. The questions are presented in Table 3; they are grouped here by their primary themes which concern domain-specific experience (Questions 1 and 2) and conflict (Questions 3 and 4), and performance (Questions 1 and 3) and working memory engagement (Questions 2 and 4). To help the reader navigate the thesis, the table also includes which papers are linked more closely with each research question.

Table 3. *Thesis questions and most relevant papers (MR) categorised by their primary themes: Domain-specific experience or conflict, and performance and working memory engagement.*

	Performance	Working Memory
Domain-Specific Experience	<p>1. Does a process' salience increase with experience and training? That is, is performance on lure items higher for experts than novices, and does it increase with training?</p> <p><i>MR: Paper 1</i></p>	<p>2. Do changes in salience affect working memory engagement? That is, does the effect of working memory constraints differ as a function of experience and training?</p> <p><i>MR: Paper 1</i></p>
Conflict	<p>3. Do changes in implicit and explicit conflict predict changes in salience across training? That is, do indicators of implicit or explicit conflict predict changes in performance?</p> <p><i>MR: Papers 2, 3, and 4</i></p>	<p>4. Do changes in implicit and explicit conflict predict changes in working memory engagement? That is, do indicators of implicit or explicit conflict predict changes in the effect of cognitive constraints?</p> <p><i>MR: Paper 2 and Paper 4</i></p>

In response to the findings emerging throughout the project, an additional fifth question was posed: Are implicit and explicit indicators of conflict signalling separate phenomena? That is, are implicit indicators of conflict related to explicit indicators of conflict, and do they both predict working memory engagement? This is discussed in more detail in the general discussion.

Methods and Measures

To investigate the research questions put forward in the previous section, six studies were conducted. This section gives an outline of the reasons for selecting the main paradigms

and measures employed in the six studies. A more in-depth discussion of the techniques pertaining to each of the studies specifically is offered in the empirical papers that follow.

The cognitive reflection test. All studies employed the CRT or variants of the CRT items (Frederick, 2005; see Appendix). The CRT was used for several reasons. Firstly, it is a quintessential bias task that has been used extensively in reasoning research, therefore, it is an appropriate starting point for bringing a novel approach to the context of reasoning. Employing a prominent bias task also facilitates an intimate discussion of the relationship between the work in this thesis and recent studies that have employed the CRT (e.g., Bago, Raelison, & De Neys, 2019b; Szollosi, Bago, Szaszi, & Aczel, 2017; Travers et al., 2016). Second, the findings from recent studies employing the CRT have been inconsistent, particularly those focused on conflict detection (e.g., compare Bago et al., 2019b; Mata et al., 2017; Travers et al., 2016). The current work goes some distance in exploring those differences and presents potential explanations for the discrepant findings.

Cognitive constraint techniques. The studies contained in the empirical papers all employ a cognitive constraint task to examine working memory engagement: either a matrix memory task load or a two-response paradigm. These techniques were employed because they have been effective in testing the engagement of working memory on bias tasks (Johnson et al., 2014; Thompson et al., 2013b; for a review De Neys, 2017b). The use of both techniques was particularly important because they allow for an examination of their convergent validity as tools for the examination of working memory engagement.

Conflict measurement techniques. Conflict was measured using confidence ratings (Papers 2, 3 and 4) and eye-tracking (Paper 3 and 4). Confidence ratings have been used to assess conflict across a multitude of bias-task studies including the CRT, as such, it is now a well-validated technique to assess conflict (e.g., De Neys et al., 2011; Hoover & Healy, 2019). In contrast, eye-tracking has not been used to test conflict detection directly. However, as noted above, Mata et al.'s (2017) eye-tracking findings could be interpreted as

an indication of conflict. They did not observe conflict effects on the bat and ball problem, but, as mentioned earlier, this null finding was small and both the authors and commentators called for more evidence (Frey et al., 2017; Mata et al., 2017). In more direct studies of conflict detection, similar techniques have been successful in assessing conflict.

In a direct test of conflict detection on base-rate neglect tasks, De Neys and Glumicic (2008) found that participants were more likely to review the critical base-rate information on lure than no lure items. De Neys and Glumicic (2008) employed a computer-based task in which the participants could review the critical base-rate information by holding down a key; it remained visible only when the key was pressed. Therefore, they could assess the duration of time a participant spent reviewing the base-rate information. They found that participants reviewed base-rate information on lure items for which the heuristic response was given more than for no lure items for which the correct response was given. Interestingly, the authors also make a distinction between measures of conflict that are explicit, such as confidence ratings, and those that are implicit, such as review effects; this distinction is an important theme of the later papers of the thesis. Eye-tracking was employed to measure implicit conflict as it offers a sensitive measure of attentional shifts, and it can be administered without altering the task's properties.

Domain-specific experience manipulations. A manipulation of domain-specific experience is included in all studies in the thesis. The first study employed a between-subjects variable for domain-specific experience using pre-existing differences in mathematical expertise between undergraduate students and mathematicians. The subsequent five studies used within-subject manipulations of domain-specific experience. That is, domain-specific experience was operationalised through a computer-based training program in which participants were given CRT-like questions and feedback and guidance

on their responses. The initial study offers an ecologically valid examination of the relationship between expertise and automatisisation, known as the “automatisisation assumption” (De Neys & Pennycook, 2019). However, the later studies, as within-subject designs, offer a more rigorous examination of the possible intra-individual factors, such as conflict, that may be involved in the engagement of working memory.

It should be noted here that the expertise garnered in short training paradigms (around 45 minutes) is not expected to reflect the changes in processing that occur over a career of mathematical training. This is why the term “domain-specific experience” is used, rather than ‘expertise’. While it would be interesting to examine the findings presented in the empirical papers with longitudinal tests of changes in domain-specific experience and expertise, the use of a short training paradigm was sufficient to elicit the changes in reasoning necessary for the examination of the research questions stated in the section above.

Structure of the Thesis

The empirical sections (Paper 1 to Paper 4) of the thesis each contain a single paper prepared for submission to academic journals. Papers 1 and 2 contain two studies each; Papers 3 and 4 contain a single study each. Each paper represents a sequential step in the progression of this project but can be read independently.

References

- Bago, B., & De Neys, W. (2017a). 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. (2017b). Rise and fall of conflicting intuitions during reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, (39), 87–92.
- Bago, B., & De Neys, W. (2019). Advancing the specification of dual process models of higher cognition: a critical test of the hybrid model view. *Thinking & Reasoning*. [Hove, East Sussex] : <https://doi.org/10.1080/13546783.2018.1552194>
- Bago, B., Raelison, M., & De Neys, W. (2019a). 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>
- Bago, B., Raelison, M., & De Neys, W. (2019b). 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>
- Benjamin, A. S., & Bjork, R. A. (2014). Retrieval fluency as a metacognitive index. In *Implicit memory and metacognition* (pp. 321–350). Psychology Press.
- De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. Retrieved from <http://pps.sagepub.com>
- 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. (2017a). Bias, conflict, and fast logic: Towards a hybrid dual process future? In W. De Neys (Ed.), *Dual Process Theory 2.0* (pp. 47–65). Abingdon, Oxon.
<https://doi.org/10.4324/9781315204550>
- De Neys, W. (2017b). *Dual process theory 2.0. 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., Moyens, E., & Ansteenwegen, D. V. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective and Behavioral Neuroscience*, 10(2), 208–216. <https://doi.org/10.3758/CABN.10.2.208>

De Neys, W., & Vanderputte, K. (2011). When less is not always more: Stereotype knowledge and reasoning development. *Developmental Psychology*, 47(2), 432–441. <https://doi.org/10.1037/a0021313>

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science*. New York, NY : <https://doi.org/10.1111/j.1467-9280.2008.02113.x>

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, 096372141985565. <https://doi.org/10.1177/0963721419855658>

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, 49(8), 709–724. <https://doi.org/10.1037/0003-066X.49.8.709>

Evans, J. S. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, 75(4), 451–468. <https://doi.org/10.1111/j.2044-8295.1984.tb01915.x>

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences*. *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

- Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review*, 13(3), 378–395.
<https://doi.org/10.3758/BF03193858>
- Evans, J. S. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*. [Hove, East Sussex] : <https://doi.org/10.1080/13546780542000005>
- 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>
- Frey, D., Bago, B., & De Neys, W. (2017). Commentary: Seeing the conflict: an attentional account of reasoning errors. *Frontiers in Psychology*, 8, 1284.
<https://doi.org/10.3389/fpsyg.2017.01284>
- Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology* (2006), 71(5), 1188–1208. <https://doi.org/10.1080/17470218.2017.1313283>
- 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>
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *The Psychological Review*. [Washington, D.C.] :
<https://doi.org/10.1037/0033-295X.103.3.592>
- Handley, S. J., & Trippas, D. (2015). Dual Processes and the Interplay between Knowledge and Structure: A New Parallel Processing Model. *The Psychology of Learning and Motivation*. San Diego : <https://doi.org/10.1016/bs.plm.2014.09.002>

- Hertwig, R., & Gigerenzer, G. (1999). The ‘conjunction fallacy’ revisited: how intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*.
[https://doi.org/10.1002/\(SICI\)1099-0771\(199912\)12:4<275::AID-BDM323>3.0.CO](https://doi.org/10.1002/(SICI)1099-0771(199912)12:4<275::AID-BDM323>3.0.CO)
- Hoover, J. D., & Healy, A. F. (2019). The bat-and-ball problem: Stronger evidence in support of a conscious error process. *Decision*. Washington, DC :
<https://doi.org/10.1037/dec0000107>
- Jacobs, J. E., & Potenza, M. (1991). The Use of Judgement Heuristics to Make Social and Object Decisions: A Developmental Perspective. *Child Development*., 62(1), 166–178. <https://doi.org/10.1111/j.1467-8624.1991.tb01522.x>
- Johnson, E. D., Tubau, E., & De Neys, W. (2014). The unbearable burden of executive load on cognitive reflection: A validation of dual process theory The unbearable burden of executive load on cognitive reflection: A validation of dual process theory. *Proceedings of the Annual Meeting of the Cognitive Science Society*, (36), 36.
- Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, 164, 56–64.
<https://doi.org/10.1016/j.actpsy.2015.12.008>
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
- Keren, G., & Schul, Y. (2009). Two Is Not Always Better Than One. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*. Malden, MA : <https://doi.org/10.1111/j.1745-6924.2009.01164.x>
- Keren, G. (2013). A Tale of Two Systems: A Scientific Advance or a Theoretical Stone Soup? Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, 8(3), 257–262. <https://doi.org/10.1177/1745691613483474>
- Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *The Psychological Review*. [Washington, D.C.] :
<https://doi.org/10.1037/a0020762>

- Mata, A., & Ferreira, M. B. (2018). Response: Commentary: Seeing the conflict: an attentional account of reasoning errors. *Frontiers in Psychology*. Pully, Switzerland : <https://doi.org/10.3389/fpsyg.2018.00024>
- Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: an attentional account of reasoning errors. *Psychonomic Bulletin and Review*, 24(6), 1980–1986. <https://doi.org/10.3758/s13423-017-1234-7>
- Mata, A., Schubert, A.-L., & Ferreira, M. B. (2014). The role of language comprehension in reasoning: How “good-enough” representations induce biases. *Cognition*. Amsterdam, etc. : <https://doi.org/10.1016/j.cognition.2014.07.011>
- Nadel, L. (2003). *Encyclopedia of cognitive science / editor-in-chief, Lynn Nadel*. London: London : Nature Pub. Group.
- Obersteiner, A., Hoof, J. Van, Verschaffel, L., & Dooren, W. Van. (2016). Who can escape the natural number bias in rational number tasks? A study involving students and experts. *British Journal of Psychology*., 107(3), 537–555. <https://doi.org/10.1111/bjop.12161>
- Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, 11(6), 988–1010. <https://doi.org/10.3758/BF03196730>
- Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, 124(1), 101–106. <https://doi.org/10.1016/j.cognition.2012.04.004>
- 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>
- Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory and Cognition*. <https://doi.org/10.3758/BF03193584>
- Simon, G., Lubin, A., Houdé, O., & De Neys, W. (2015). Anterior cingulate cortex and

intuitive bias detection during number conservation. *Cognitive Neuroscience*. Hove, East Sussex : <https://doi.org/10.1080/17588928.2015.1036847>

Sloman, S. (2014). Two systems of reasoning: An update. In *Dual-process theories of the social mind*. (pp. 69–79). New York, NY, US: The Guilford Press.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1), 3–22. <https://doi.org/10.1037/0033-2909.119.1.3>

Svedholm-Häkkinen, A. M. (2015). Highly reflective reasoners show no signs of belief inhibition. *Acta Psychologica*. The Hague : <https://doi.org/10.1016/j.actpsy.2014.11.008>

Szollosi, A., Bago, B., Szaszi, B., & Aczel, B. (2017). Exploring the determinants of confidence in the bat-and-ball problem. *Acta Psychologica*, 180, 1–7. <https://doi.org/10.1016/j.actpsy.2017.08.003>

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., Striener, C. L., Reikoff, R., Gunter, R. W., & Campbell, J. I. D. (2003). Syllogistic reasoning time: Disconfirmation disconfirmed. *Psychonomic Bulletin & Review*. Austin, TX : <https://doi.org/10.3758/BF03196483>

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013a). 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>

Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013b). 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>

- Toplak, Maggie E, West, Richard F, & Stanovich, Keith E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, 20(2), 147-168. <https://doi.org/10.1080/13546783.2013.844729>
- 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>
- Trippas, D., & Handley, S. J. (2017). The parallel processing model of belief bias: Review and extensions. In W. De Neys (Ed.), *Dual process theory 2.0* (1st ed.). London: Routledge. <https://doi.org/10.4324/9781315204550>

Paper 1

Domain-Specific Experience and Dual-Process Thinking

Zoe A. Purcell, Colin A. Wastell and Naomi Sweller

Department of Psychology, Macquarie University, Sydney

This paper has been submitted for publication to

Thinking & Reasoning

Abstract

Prominent dual process models assert that reasoning processes can transition from effortful (i.e. Type 2) to intuitive (i.e. Type 1) with increases in domain-specific experience (DSE). In two studies we directly examine this assertion. We test the relationship between DSE, operationalised via mathematical experience, and performance on the cognitive reflection test (CRT; Frederick, 2005). Additionally, we employ cognitive constraint paradigms in which participants have to complete a task of varying complexity while concurrently completing the CRT. In Study 1, we manipulate DSE via pre-existing differences in mathematical experience; demonstrating changes in thinking Type across real-world differences in DSE. In Study 2, we manipulate DSE via a mathematical training paradigm. Our main findings suggest that a parabolic relationship exists between DSE and thinking Types such that low DSE is associated with Type 1 processing, intermediate DSE is associated with Type 2 processing, and high DSE is associated with Type 1 processing. We relate these findings to the dual process literature and suggest that DSE may account for previously inconsistent findings.

Domain-Specific Experience and Dual-Process Thinking

A novel problem or task may seem difficult at first, but with enough practice, it can become easy and routine. Take, for example, the experience of learning to drive a car. At first it can seem overwhelming. To start, you must learn the many functions of the equipment: the clutch, the accelerator, the gear-stick, the indicators, and the brake(!). Then you must learn how to use these—in synchrony—to suit the conditions around you, maintaining a safe distance from the car in front of you, staying in the correct lane, and paying attention to road signs. Eventually, after many hours of practice and experience, these skills can become intuitive and seamlessly integrated, so much so that you may even be able to hold a conversation at the same time. Practice and the process of learning, be it in the domain of driving or mathematics, is often accompanied by some mild cognitive unease and effortful thinking, but—over time—can eventuate in a transition from effortful to effortless thinking.

Reasoning and thinking scholars, particularly dual process theorists, are interested in the differences between effortful (Type-2) and intuitive (Type-1) thinking, and have suggested that some Type-1 processes may be the product of Type-2 processes having been practiced to the point of automation (e.g., Evans & Stanovich, 2013; Kahneman, 2011). However, the suggestion that Type-2 processes can become Type-1 processes with increases in DSE, like driving or mathematics, has not been explicitly tested. Reasoning theorists assert that the core distinction between thinking Types is based on working memory (e.g., Evans & Stanovich, 2013). Working memory is thought to be a relatively stable “hardware” of an individual’s higher cognition, used to hold information that can be accessed and manipulated for a short time, but which is vulnerable to interference from competing cognitive tasks (e.g., Baddeley, 1986; Engle, Tuholski, Laughlin, & Conway, 1999; Hambrick & Engle, 2002). Fittingly, the working memory literature includes models and research about expertise and learning that are pertinent to the current

investigation into DSE and thinking Types. Hence, the current article first explores the literature from the *reasoning* and *working memory* fields. The key aspects of this literature are then integrated in an empirical examination of the relationship between DSE and thinking Types.

Reasoning and the Dual Process Model

The key tenet of dual process theory is that reasoning is achieved via two distinct types of processes: Type 1 which is automatic and intuitive, and Type 2 which is deliberate and effortful (e.g., Epstein, 1994; Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996). Evans and Stanovich (2013) defined Type 1 processes as autonomous: they do not require controlled attention, are not dependent on input from high-level control systems, and—importantly—do not require working memory. Type 2 processes are characterised by the engagement of a general-purpose system, are responsible for cognitive decoupling and hypothetical thinking, and—in contrast to Type 1 processes—require working memory.

Dual process theories acknowledge that the Types of thinking may interact (for a review, see De Neys, 2017). While the hypothesised nature of this interaction differs between models on some aspects, there is a consensus that Type-2 processes can transition into Type-1 processes with practice (e.g., Epstein, 1994; Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996). This transition phenomenon (from Type-2 to Type-1), also known as the process of automation, is more thoroughly developed in theories such as Klein's (2003) naturalistic decision-making theory, Gigerenzer's (2007) fast and frugal heuristics theory, Reyna's (2012) fuzzy trace theory, and Wastell's (2014) complex emergent theory. However, it remains under-developed within dual process models.

The role of DSE and automation in thinking can be incorporated by some dual process models more easily than by others. Particularly suitable, are those models suggesting that multiple Type-1 processes can be triggered at the outset of problem solving (De Neys, 2012; Handley & Trippas, 2015; Pennycook, Fugelsang, & Koehler, 2015;

Trippas & Handley, 2017). Although many models could accommodate the automation phenomenon with small modifications or clarifications, the logical intuitions model is arguably best suited for the task. The logical intuition model asserts that, when faced with a problem, multiple Type-1 reasoning processes may be activated (De Neys, 2012, 2014). The process with the highest activation is actioned or “wins out.” If two or more processes have similar levels of activation, conflict may occur which can manifest as a sense of cognitive unease or uncertainty (Bago & De Neys, 2017; De Neys, 2012, 2014). This conflict, the model suggests, may be involved in the engagement of Type 2—working memory dependent—processes. The consideration of competing processes is particularly useful when exploring how processes might be influenced by DSE.

The more practice a certain process has received, the more likely it is to be actioned in the future. In terms of the logical intuition model, the increased likelihood of a process winning the activation competition may be conceptualised as an increase in activation potential. Figure 1 shows the hypothetical activation of two processes, one leading to a heuristic response, the other leading to a logical response. Here, the reasoner is being trained to use the logical process. At Time 1, prior to training, the reasoner gives the heuristic response, experiences little conflict, and subsequently engages very little working memory (i.e. Type-1). After some training, at Times 2 and 3, the activation of the logical process comes closer to that of the heuristic response. Hence, the reasoner may experience more conflict and greater working memory engagement (i.e. Type-2). At Times 3 and 4, the reasoner gives the correct, logical response. From Time 3 to Time 4, the difference between activations becomes larger and the reasoner experiences less conflict—decreasing the working memory engagement (i.e. Type-1). Although this example is highly simplified, it demonstrates the principle of transitioning from heuristic to logical processes via training, and the parabolic relationship between thinking Type and increases in DSE.

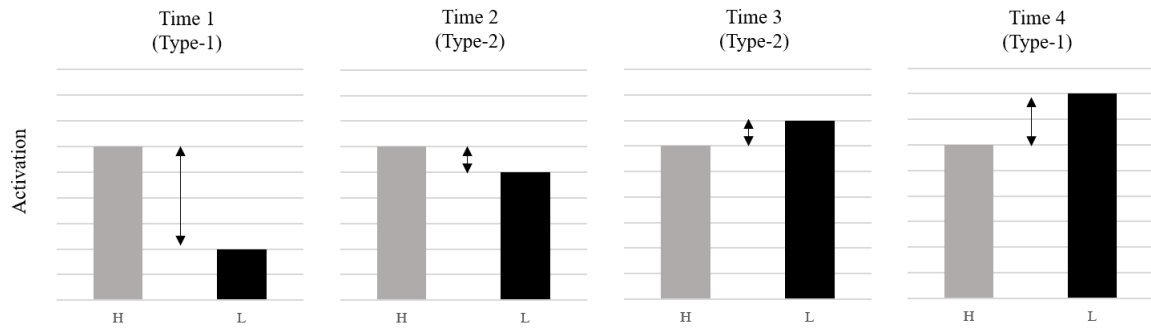


Figure 1. *Conjectural model of the increasing activation potential of a logical process across training and the interaction with thinking Type. H=Heuristic Process, L=Logical Process.*

Studies on reasoning often require participants to solve ‘bias tasks’ while under cognitive constraints, such as a timing deadline or an additional cognitive load, to determine whether working memory was engaged (e.g., Bago & De Neys, 2019). Bias tasks are thought to generate responses that can be clearly distinguished as the result of Type-1 or Type-2 thinking. A quintessential bias task is the cognitive reflection test (CRT; Frederick, 2005), a three-item test that commonly elicits intuitive, erroneous responses (Frederick, 2005). Consider the first CRT problem: “A bat and a ball together cost \$1.10. The bat costs \$1 more than the ball. How much does the ball cost?” This problem often prompts the incorrect response—10c, despite most responders having the ability to reach the correct solution—5c. As the name suggests, the CRT was originally proposed as a test of reflective thinking. Like many bias tasks, correct responses have been interpreted as the result of Type-2 processes intervening and correcting the default Type-1 solution (e.g., Kahneman, 2011; Kahneman & Frederick, 2005; Toplak, West, & Stanovich, 2011). Recently, studies have employed cognitive constraints to test this ‘corrective interpretation’ (e.g., Bago & De Neys, 2017, 2019). However, those examining the effects of cognitive constraints on the CRT have yielded inconsistent results.

Dual process theorists assert that Type 2 processes demand more time and cognitive resources than Type 1 processes (Kahneman, 2011; Kahneman & Frederick, 2005). Cognitive constraints that deprive the reasoner of time or working memory

resources have, therefore, been used to determine the thinking Type employed by participants during bias tasks (Bago & De Neys, 2017). Some studies have found support for the corrective interpretation of CRT findings. For example, Johnson, Tubau and De Neys (2014, 2016) employed a visuospatial memory task as a cognitive constraint. Participants completed CRT-like problems under one of three conditions: no-load, low load, or high load. Those in the load conditions were required to complete the CRT while remembering a 3x3 dot matrix pattern, which has been shown to interfere with executive resources (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). In line with the corrective interpretation of CRT responses, they found that participants' performance on the CRT fell when they were required to perform a simultaneous, cognitively-taxing task.

In contrast, Bago and De Neys (2019) found that most participants giving correct responses on the bat-and-ball problem (the first item in the CRT) were able to do so under conditions of cognitive constraint. They employed a two-response paradigm in which participants were given two attempts to complete the bat-and-ball problem; the first attempt under cognitive load and time-pressure, and the second, without cognitive constraints. They found that most respondents who gave a correct answer at the second attempt, also gave the correct answer for their first attempt. In their studies that employed a free-response style (i.e., not multiple choice), the correct answer was given at both attempts on 20% of trials, compared to 7% of trials in which participants changed from an initially incorrect response at attempt one, to a correct response at attempt two (Studies 4 and 5). These findings indicate that most respondents who gave a correct answer could do so under cognitive restrictions, but also that a substantial portion of the participants were able to change their initially incorrect response to a correct response at the second attempt—that is, only without cognitive restrictions.

Bago and De Neys (2019) considered that the correct responses provided under cognitive restriction may be due to those reasoners having automatised the required

process. They acknowledge that automation is the goal of many learning contexts and that years of exposure during high school or other mathematical training may have helped these reasoners to do just that. In line with this interpretation, we note that the portion who did require Type-2 conditions to successfully complete the problem may have been at an intermediate point of automation, where they still needed to go through a process of deliberation that required cognitive resources. In other words, these participants may have been at a point of learning or DSE where cognitive resources were required to reach the correct solution. Previous studies that have tested the effect of cognitive constraints on CRT performance have yielded inconsistent findings (Bago & De Neys, 2019; Johnson, Tubeau & De Neys, 2014, 2016). In accordance with the working memory literature, the studies in the current article examine whether previous inconsistencies might be accounted for, at least in part, by mathematical DSE.

Working Memory

The pattern of working memory engagement proposed in the dual process logical intuition model is complementary to literature on working memory that suggests increased DSE is associated with decreased dependency on working memory resources (e.g., Ericsson & Kintsch, 1995; Gobet & Simon, 1996; Guida, Gobet, Tardieu, & Nicolas, 2012). Ericsson and Kintsch's (1995) long-term working memory model (LT-WM) proposes that the units of information typically stored in working memory can be efficiently stored in, and retrieved from, long-term memory by skilled performers. Building on Chase and Ericsson's (1982) skilled memory theory, the LT-WM model suggests that high levels of domain-knowledge developed through DSE allows the individual to encode information in long-term memory with advanced storage and retrieval structures. These structures facilitate rapid storage and retrieval of information from long-term memory. The skilled individual can therefore use long-term memory storage for domain-specific units of information in a temporary manner, such that they can, to some extent, circumvent the

limitations of their working memory capacity (WMC).

This temporary long-term memory storage differs from temporary working memory storage and short-term memory storage (Ericsson & Kintsch, 1995). Short term memory does not facilitate the combination or manipulation of information—as in working memory, nor does it offer protection from the detrimental effects of cognitive constraints—as in long-term memory (Ericsson & Kintsch, 1995). The LT-WM model is supported by studies demonstrating that skilled activities, such as piano playing (Allport, Antonis, & Reynolds, 1972) and reading (Glanzer, Dorfman, & Kaplan, 1981; Glanzer, Fischer, & Dorfman, 1984), are the least impaired by concurrent tasks. Should increases in DSE lead to more efficient retrieval and storage structures in long-term memory, it follows that increases in DSE will alleviate working memory resources, in turn, allowing greater concurrent combinations of information. The greater the potential for concurrent combination, the greater the potential for learning.

Learning was linked to concurrent combination by Hebb (1949), who suggested that a connection between two or more ideas can be formed when their representations are simultaneously activated. Similarly, Anderson's (1983) ACT-R model proposed that connections develop through the concurrent representation of events. Concurrent activation is viewed as a primary function of working memory (Hambrick & Engle, 2003). The link between learning and working memory is supported by studies demonstrating a positive relationship between WMC and complex learning (e.g., Kyllonen & Dennis, 1996; Kyllonen & L. Stephens, 1990). These developments in the learning and working memory literature have important implications for the use of DSE manipulations in reasoning studies.

Reasoning studies may employ DSE manipulations via between-subjects designs. For example, it may be of interest to the researcher to compare the reasoning characteristics of novices, intermediates and experts in a particular domain. This between-

subjects design should reveal differences in working memory dependence for the successful completion of domain-relevant tasks. Participants with very little domain-specific knowledge (i.e., novices with low DSE) are more likely to provide incorrect Type-1 responses and show little conflict or awareness of the error. Whereas, participants at intermediate levels of DSE may be more dependent on Type-2 working memory dependent processes. Once a reasoner can reach the correct response, higher DSE should lead to less dependence on working memory. In line with Hambrick and Oswald (2005), a between-subject DSE design may yield independent effects of WMC and DSE on performance on domain-relevant tasks.

In contrast, within-subject designs may produce interactive effects between WMC and DSE on task performance. People with higher working memory capacities learn more quickly than those with lower working memory capacities (Kyllonen & Dennis, 1996; Kyllonen & Stephens, 1990). Therefore, studies that operationalise DSE via training may generate cumulative WMC effects in which learning facilitates LT-WM effects which, in turn, “frees up” working memory space allowing for even more concurrent combinations and further learning. The working memory and learning literature has different implications for between-subject DSE manipulations operationalised through pre-existing DSE levels (as in Study 1 below), and within-subject DSE manipulations operationalised through training (as in Study 2 below). These relationships between DSE, working memory dependence, and task performance can be examined by combining DSE manipulations using cognitive constraints.

The Current Studies

In two studies, we establish a trajectory of deliberation across changes in DSE. First, in a real-world between-subjects DSE design and, second, in a lab-based within-subject DSE design. This is the first instance that we know of where DSE manipulations have been used in conjunction with the CRT and cognitive constraints. We chose to focus

on the CRT for several reasons. First, it is a mathematical problem-solving task and has been shown to correlate with numeracy (e.g., Cokely & Kelley, 2009; Liberali, Reyna, Furlan, Stein, & Pardo, 2012). It therefore lends itself to the straightforward DSE manipulation via mathematical experience. Second, most people give the incorrect response on the CRT, leaving a greater capacity for improved performance (Frederick, 2005). The few people giving correct responses on the CRT have done so with variable dependence on working memory which may, in part, be accounted for by DSE (Bago & De Neys, 2019; Johnson et al., 2014, 2016). Third, while DSE manipulations might be applied to many tasks, the CRT is a centrepiece of prominent reasoning models and features heavily in dual process literature. The CRT is, therefore, an appropriate place to start when first considering the potential empirical contribution of DSE to reasoning research.

In both studies we examined the relationship between DSE, cognitive constraint, and performance on the CRT. Cognitive constraint was manipulated in both studies through the use of dot matrix memory tasks (e.g., Bago & De Neys, 2019). Study 1 employed between-subjects manipulation of DSE in a 2 (cognitive constraint) x 3 (DSE) design. Study 2 employed a within-subject manipulation of DSE in a 3 (cognitive constraint) x 3 (test point) design. The hypotheses for each study are outlined in the respective sections that follow.

Study 1

In Study 1, DSE was incorporated as a pseudo-independent variable—participants were classified according to their university course or occupation; CRT performance was compared between participants with low, intermediate, and high mathematical experience. Cognitive constraint was manipulated by placing participants into load or no-load conditions. We expected that, for CRT performance:

- 1) Participants with more DSE would outperform those with less DSE;
- 2) Participants would perform better when no-load was imposed than if a

load was imposed;

3) The performance of participants with intermediate DSE would be detrimentally affected by the load to a greater extent than for those with low or high DSE.

Method

Participants and Design. Study 1 employed a 2 (cognitive constraint) x 3 (DSE) between-subjects design. Only one female participant qualified for the high DSE group, therefore, this participant was excluded and we used an all male sample for the low and intermediate DSE groups¹. Final participants were 65 males, with ages ranging from 18 to 72 ($M=25.46$ years, $SD=12.96$). Participants were randomly allocated to no-load ($N=34$) or load ($N=31$) conditions. Low DSE participants were 26 undergraduate psychology students at Macquarie University, Sydney (aged $M=21.96$, $SD=10.65$). Intermediate DSE participants were 24 undergraduates in actuarial, science, or engineering courses at Macquarie University, Sydney (aged $M=20.33$, $SD=4.00$). High DSE participants were 15 postgraduate mathematical students at Macquarie University, Sydney, or professional mathematicians (aged $M=39.73$, $SD=15.94$). Undergraduate participants were recruited from Macquarie University, and non-students were recruited via a mathematics website. Psychology students were awarded course credit for participation. Non-psychology students were offered the chance to enter a draw for one of three \$AU50 vouchers.

Materials

CRT. The three-item CRT was used (Frederick, 2005). There was no time limit to solve the problems, and a free-response format was used (i.e. not multiple choice). Participants entered a number in the units specified on screen. Total scores ranged from 0-3.

¹ Gender has been shown to affect CRT performance through mathematical anxiety (e.g., Frederick, 2005; see Morsanyi, Prado & Richland, 2018, and Primi et al., 2018).

Matrix memory task. Matrices were used as cognitive constraints. Participants were required to memorise 3x3 grids with four coloured squares presented for 800ms (see Figure 2). This task was adapted from previous dot matrix memory tasks for use on the Qualtrics survey platform. The matrices had four coloured squares that formed “three-piece” patterns (see Figure 2; Bethell-Fox & Shepard, 1988; Verschueren, Schaeken, & d’Ydewall, 2004). Matrices have been used effectively to impose secondary loads on reasoning processes (e.g., Bago & De Neys, 2019; Johnson et al., 2014, 2016). Matrix performance scores were calculated by scoring each coloured (or not) square as correct or incorrect. Scores are reported as percentages.

Numeracy. The Problem-Solving Test (PST) was used to substantiate the categorisation of participants by course and occupation as a reflection of DSE. The PST is a measure of numeracy containing 12 mathematical problems (Hegarty, Mayer, & Monk, 1995). This scale was selected because it contains items that are similar to those in the CRT; they are presented as written problems and contain similar quantities, units and calculations. There were no time limits imposed. An example of a PST item is: “At McDonald's, workers earn \$6.00 per hour. This is 50 cents less per hour than workers at Wendy's. If you work for 8 hours, how much will you earn at Wendy's?” Scores could range from 0-12.

Procedure. Study 1 was completed online via Qualtrics. After consent was obtained, participants completed a series of demographic questions. They were then given instructions for the general procedure. Further instructions were presented on each page as appropriate, for example: “Submit your final answer only. Use numbers only (up to 2 decimal places). Exclude symbols or words e.g. \$, cents, km.” Participants then completed the CRT and matrix memory tasks. The order of the question and matrix presentation was varied between conditions such that the no-load group solved the matrix task and CRT problems separately whereas the load group were required to solve the CRT problems

while maintaining the grid pattern in working memory (see Figure 2). Those in the load condition were therefore required to solve the problems under cognitive constraint, whereas those in the no-load condition experienced no cognitive constraint. The information component of the questions was presented first to minimise the effect of the load on the comprehension process (Van Lier, Revlin, & De Neys, 2013). Finally, participants completed the PST.

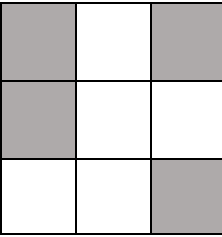
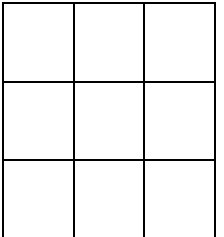
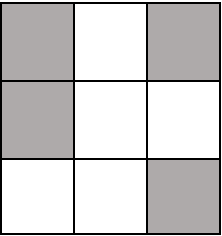
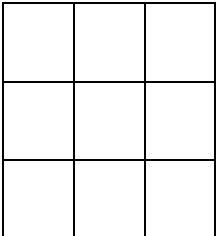
Presentation Order	First	Second	Third	Fourth
(a) No-load Condition	“A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball.”	Memorise: 	Recall: 	“A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost?”
(b) Load Condition	“A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball.”	Memorise: 	“A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost?”	Recall: 

Figure 2. Example of procedure for CRT problem 1 for (a) no-load and (b) load conditions. Each of the four elements were presented separately. The order of the elements differed between conditions.

Results

Preliminary analysis. Performance on the PST was analysed to check the categorisation of mathematical DSE by course and occupation. High DSE participants scored higher ($M=10.27$, $SD=1.53$) than intermediate DSE participants ($M=9.70$, $SD=1.79$) who, in turn, scored higher than low DSE participants ($M=9.35$, $SD=1.94$). These differences were not significant, $F(2, 61)=1.24$, $p=.295$, $\eta^2_p=.04$, likely due to ceiling effects. Although the trend was positive, numeracy and performance on the CRT were not significantly associated, $r(64)=.246$, $p=.056$; again, this is likely due to ceiling effects. However, the positive trend between DSE and PST scores supports the categorisation of participants by course and occupation as a measure of mathematical experience.

Additionally, that all participants performed well on the PST suggests that differences in CRT performance were not due to a general lack of mathematical ability by any group.

Performance on the matrix memory tasks was analysed to check for systematic differences in task preference between the DSE groups; that is, whether participants prioritised the matrix task over the CRT task, or vice versa. The effect of DSE and load condition on matrix performance was assessed using a two-way ANOVA. Participants in the no-load condition ($M=92.05$, $SD=14.39$) outperformed those in the load condition ($M=80.52$, $SD=11.67$) on the matrix memory task, $F(1, 59)=16.16$, $p<.001$, $n^2_p=.22$. The main effect of DSE on matrix performance was not significant, $F(2, 59)=.74$, $p=.480$, $n^2_p=.03$. There was a significant interaction between DSE and condition on matrix performance: $F(2, 59)=3.29$, $p=.044$, $n^2_p=.10$. This indicated that the effect of load condition on matrix performance was different at different levels of DSE. An interaction contrast test was run comparing matrix performance for those in the low- and high-DSE groups combined, compared to those in the intermediate DSE group. This contrast revealed that the effect of load condition on matrix performance was greater for intermediate DSE participants than low and high DSE participants, $F(1, 59)=6.52$, $p=.013$, $n^2_p=.10$. This indicated that intermediate DSE participants may have prioritised the CRT over the matrix task to a greater extent than did low and high DSE participants. Conversely, low and high DSE participants may have prioritised the matrix task over the CRT more than intermediate DSE participants. This pattern of results suggested a potential for different task preference between low/high DSE groups and the intermediate DSE group in line with the primary hypotheses. Therefore, we included matrix performance as a potential covariate in the main analyses.

Main analysis. To test our three hypotheses, we used a 2 (load) x 3 (DSE) between-subjects ANOVA, with pairwise comparisons to follow up main-effects, and interaction contrasts to follow up interaction effects. CRT scores were significantly higher

for participants with greater DSE, averaged across load condition, $F(2, 59)=31.02, p<.001, n^2_p=.51$. High DSE participants ($M=2.60, SD=.63$) outperformed intermediate DSE participants ($M= 1.54, SD=1.06$), $F(1, 64)=14.76, p<.001, n^2_p=.19$, who outperformed low DSE participants ($M=.65, SD=.69$) $F(1, 64)=14.04, p<.001, n^2_p=.19$. The main effect of load was not significant when averaged across DSE, $F(1, 59)=2.04, p=.158, n^2_p=.03$.

However, there was a significant interaction between DSE and condition on CRT performance, $F(2, 59)=6.66, p=.002, n^2_p=.18$. An interaction contrast test was run to compare the difference in CRT scores between the load and no-load conditions for the low and high DSE groups combined, compared to the intermediate DSE group. Results revealed that the effect of load condition was greater for intermediates than for the low and high DSE groups, $F(1, 59)=11.18, p=.001, n^2_p=.16$. These results are presented in Figure 3. Due to differences between groups on matrix performance (reported above) we also ran this analysis with matrix performance included as a potential covariate, however, the pattern of results was unchanged². Therefore, we have reported the results excluding matrix performance.

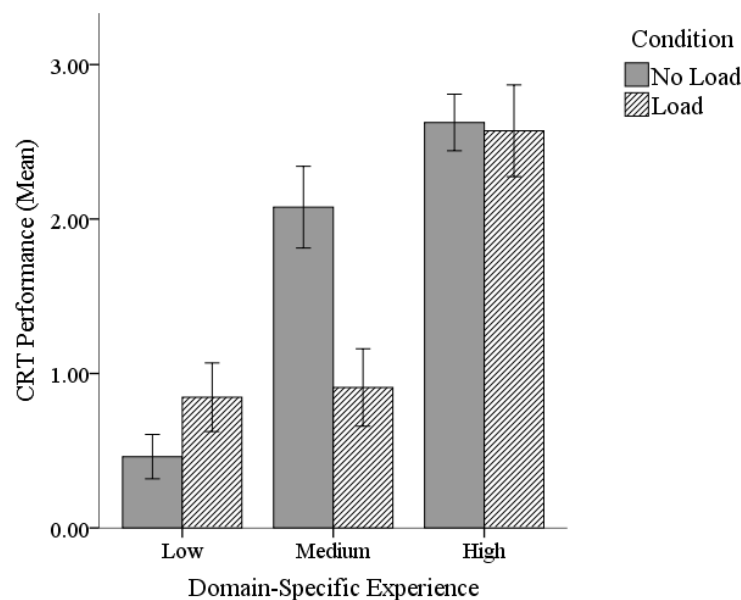


Figure 3. CRT performance by level of DSE and cognitive load. Error bars reflect +/- 1 SE.

² These results are available from the authors on request.

Discussion

Using a quasi-experimental design, Study 1 examined three hypotheses: that those with higher DSE would outperform those with lower DSE; that those in the no-load condition would outperform those in the load condition; and, that those with intermediate DSE would be affected by the load to a greater extent than for those with low or high DSE. The findings demonstrated a positive relationship between DSE and CRT performance. This finding supports our first hypothesis and is in line with previous studies that have shown positive associations between CRT performance and mathematical factors such as numeracy (Sinayev & Peters, 2015; Welsh, Burns, & Delfabbro, 2013), and SAT scores (Frederick, 2005; Obrecht, Chapman, & Gelman, 2009; Thompson et al., 2013). Our second hypothesis was not supported. The results from Study 1 did not show an effect of load when averaged across DSE. In regard to the effect of load, this is similar to Bago and De Neys' (2019) study that found that participants were able to complete the bat-and-ball problem under cognitive constraint but contrasts with Johnson et al., (2014, 2016) who observed a detrimental effect of load on CRT performance. However, as the results pertaining to our third hypothesis indicate, the effect of load should be considered in relation to DSE.

In support of our third hypothesis, load and DSE interacted to affect CRT performance. Participants in the intermediate DSE group were more affected by the load than participants in the low or high DSE groups. This indicates that participants in the intermediate DSE group were able to reach the correct solutions on the CRT but were more dependent on working memory than their low and high DSE counterparts. In dual process terms, the intermediate group appears to have engaged Type-2 thinking to a greater extent than the other two groups. The high DSE group, conversely, did not show lowered performance under load, suggesting they may have been using Type-1 processing to complete the CRT. Those in the low DSE group performed poorly regardless of load which

could be interpreted as an indication that they had engaged Type-1 thinking. However, this could be due to a floor effect. That is, the load may not have affected their performance because it was already so low. It is possible that this group was using Type-2 processing but failed to reach a correct solution. Hence, it is unclear which Type of processing may have been employed by the low DSE participants. However, the pattern of results for the high and intermediate DSE groups demonstrate Type-1 and Type-2 processing, respectively.

The use of real-world DSE groups increased the study's ecological validity but also introduced limitations concerning sample-size and potential confounds. Due to the inclusion of an expert, high DSE population, Study 1 had a small sample size. This highlights the need for a paradigm that does not require a restricted sample, for example, by using a within-subject training paradigm as in Study 2. By using an all male sample we prevented gender from potentially confounding our independent variables, however, age differed systematically between DSE groups. Among adults, aging is related to declines in executive functioning and could increase the detrimental impact of cognitive constraints; for working memory, these declines are more prominent amongst people in their 70s and 80s (Buckner, 2004; Park et al., 1996). In Study 1, the high DSE group were older than the low and intermediate participants. However, high DSE participants were mostly under 70 years old and were impacted by the load to a lesser extent than the younger, intermediate DSE group. While this suggests that age-based differences did not suppress the expected pattern of results in Study 1, it should be considered in future DSE studies. Using pseudo-independent variables for DSE is likely to restrict sample size, bring demographic disparities mirroring those in society, and consequently, introduce potentially confounding factors. Using a within-subject manipulation of DSE, as in Study 2, is one way to help prevent these issues.

Participants in the low, intermediate, and high DSE groups may have differed on

more than just demographic characteristics. For example, DSE groups may have differed on prior exposure to the CRT or psychological factors such as intelligence and WMC. It could be that mathematicians, because of an interest in mathematical problems, or psychology students, because of an interest in psychology, may have completed the CRT. Although, Bialek and Pennycook (2018) found that exposure alone did not impact performance on the CRT, we cannot be sure if this influenced the findings from Study 1. Moreover, while exposure to the questions alone did alter performance in Bialek and Pennycook's study, it is possible that our sample included participants who had been exposed to both the questions and the answers. One way to reduce the likelihood of exposure confounds is to couple a within-subject paradigm with random allocation to conditions, as in Study 2.

DSE groups may have differed on other factors that have demonstrated associations with CRT performance, such as thinking dispositions, intelligence or working memory (e.g., Toplak et al., 2011). However, the most important factor for consideration, given the manipulation of cognitive load in Study 1, is WMC. It could be that high DSE participants had larger working memory capacities than intermediate and low DSE participants. Consequently, the high DSE participants may have been able to engage in Type 2 processing even with the additional cognitive load. If this was the case, the load manipulation—while successfully used in previous studies—may not have been large enough to “knock out” the WMC of those in the high DSE group. Coupled with a potential ceiling effect, we cannot be sure that high-DSE participants were using Type 1 processes. This issue is linked to the theoretical frustration in that no dual process proponent has defined the threshold of working memory engagement at which we might classify a process as one Type or the other (Keren, 2013; Kruglanski, 2013; Osman, 2004, 2013). Until that point, studies such as this will face the criticism that participants succeeding despite cognitive constraint were using Type 2 processing and that the cognitive load

simply wasn't large enough. However, there are some ways to appease such critics, for example, by controlling for WMC or including a range of cognitive loads—particularly with harder constraints. Study 2 includes both a harder cognitive constraint and a measure of WMC.

Study 1 is the first (that we know of) to combine DSE and cognitive constraints in a study of CRT performance. Study 1 showed preliminary support for the dual process assertion that Type 2 processes can become Type 1 processes with increases in DSE. It also supports the logical intuition model's assertion that logico-mathematical principles can be affected via Type 1 processes (De Neys, 2012). Additionally, Study 1 paints a very promising—albeit preliminary—picture for the inclusion of DSE in future investigations that aim to elicit Type 1 and Type 2 processing. Applying the interpretation of previous cognitive-constraint studies and dual process theory (e.g., Bago & De Neys, 2019), the results from Study 1 indicated that low-DSE participants were using Type 1 processes, intermediate-DSE participants were using Type 2 processing, and high-DSE participants were using Type 1 processing.

Study 2

Study 2 addressed the same research question as in Study 1: Can Type 2 processes become Type 1 processes with increased DSE? However, a different approach was taken in order to overcome the limitations in Study 1 (i.e., controlling for potential confounds) and to develop a more sustainable method for the inclusion of DSE in future studies (i.e., removing the need for recruiting expert populations). This was achieved by including a within-subject training manipulation of DSE with random allocation to load conditions. Additionally, we included a harder cognitive load, and measures of WMC and other potentially confounding factors. This approach yielded a 3 (cognitive constraint) x 3 (test point) mixed design. As in Study 1, we manipulated cognitive constraint with matrix tasks, in this case with three levels: low, intermediate, and high. We included an DSE

manipulation via training; participants' CRT performance was measured at three test points: before training (test point 1; T1), half way through training (test point 2; T2), and after training (test point 3; T3).

As explored earlier in the article, WMC may have a *dynamic effect* on performance, due to learning and automation, as well as a *static effect*, due to the relationship between WMC and cognitive load. WMC begets learning, learning leads to automation and “frees up” WMC (LT-WM) which, in turn, begets learning. These relationships can, therefore, form a dynamic and cumulative effect of WMC on performance across training. However, there is also a static effect of WMC on performance. Holding the point of automation constant, the impact of a cognitive load may differ according to an individual's WMC. To consider both the dynamic impact of WMC through learning, as well as the static impact of WMC on cognitive load effects, let us explore a hypothetical case. Imagine a person who has a WMC of 10 units and that our load task (the matrix) takes 6 units of WMC to maintain. For this person, the load would impact their performance if the main task (the CRT) required 5 or more units of WMC. Our manipulation of DSE was expected to reduce the amount of WMC required by the CRT task. Hence at T1, the task may require 5 units, but after 1 block of training it may have required only 4 units, and after another block of training, it may have required only 3 units of WMC. For this person, with a WMC of 10 units, their WMC would be exhausted at T1, but not T2 and T3. They would be expected to be able to successfully complete the CRT at T2 and T3. However, for a participant with lower WMC, say 8 units, they would be expected to successfully complete the CRT at T3 but not at T1 and T2 because their WMC is exceeded by the combined load of the CRT and the matrix tasks. With these factors in mind we anticipated an interactive effect of WMC, test point and load on performance.

A simplified version of these hypotheses is presented in Figure 4 which shows a transition from failure to success. The first column shows hypothesised performance at T1;

participants with high WMC demonstrate higher scores with low load, but participants do not perform well overall. This is in line with previous studies that have shown a positive association between WMC and CRT performance (Stuppel, Gale, & Richmond, 2013; Toplak et al., 2011). The second column shows hypothesised performance at T2; participants with high WMC can now complete the task even under a medium load, and participants with medium WMC can complete the task under a low load. The third column shows hypothesised performance at T3; those with high WMC can complete the task under all levels of load, those with medium WMC can complete it under a medium load, and those with low WMC can complete it under a low load. At this point we should note that the extent of training needed to automate the processes required for the successful completion of the CRT was unknown, hence, we put forward these hypotheses tentatively regarding specific test points, but strongly regarding the expected pattern. In other words, we expected this pattern to emerge but were unsure at which point of training.

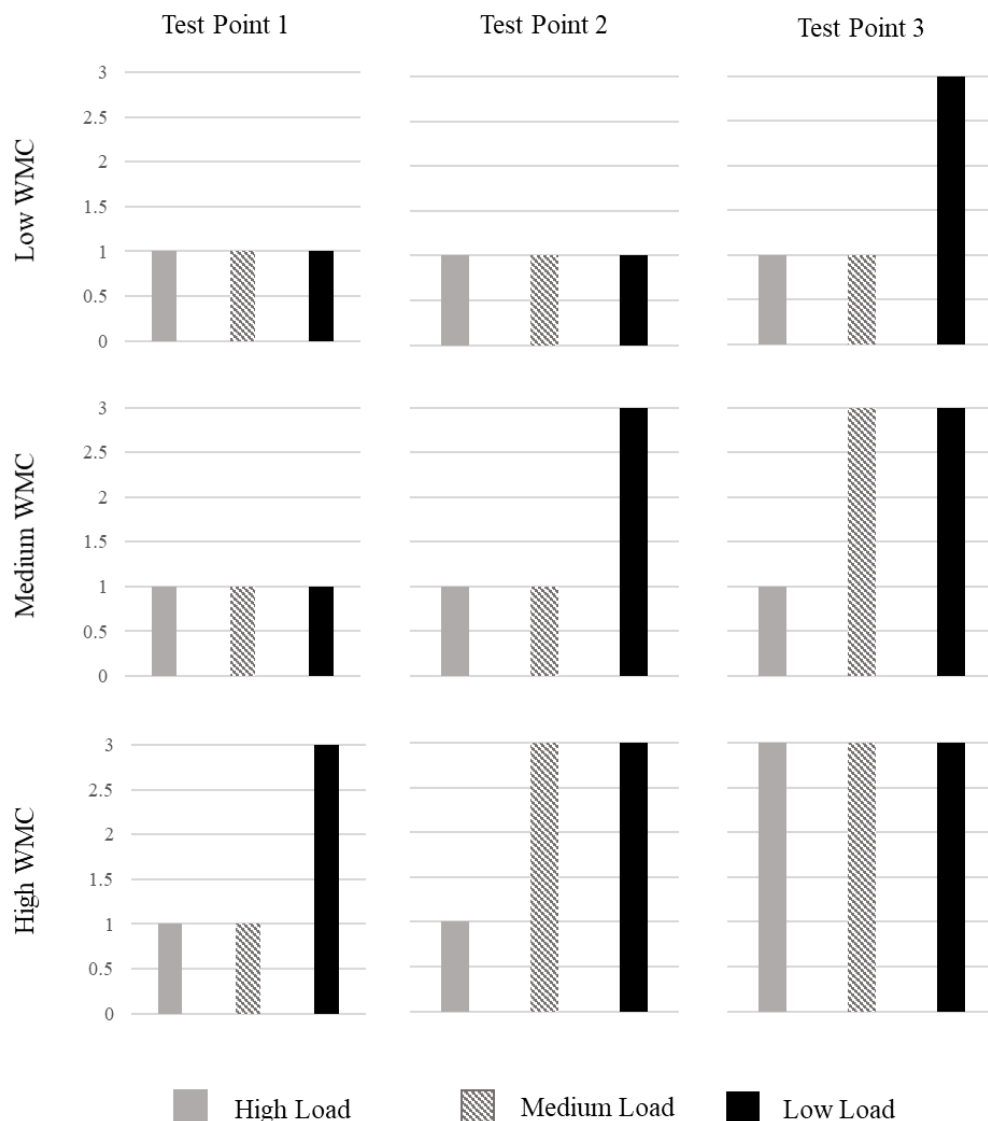


Figure 4. *Hypothesised interactive effect between WMC, load, and test point on CRT performance.*

In sum, we expected that:

- 1) Performance would increase with increases in DSE;
- 2) Training, load, and WMC would interact to affect performance such that at T2, participants with lower WMC would be negatively affected by the load manipulation to a greater extent than those with higher WMC; this difference was expected to be smaller at T1 and T3 than at T2.

Method

Participants and Design. Study 2 used a 3 (DSE: Test point) x 3 (Load) x WWC (continuous predictor) mixed design. A total of 85 participants were recruited from

undergraduate psychology at Macquarie University, Sydney (22 males, 61 females, 2 unspecified). Ages ranged from 17 to 43 ($M=20.28$, $SD=4.00$). DSE was included as a within-subject factor with three levels: T1, T2, and T3. Load was included as a between-subjects factor and participants were randomly allocated to one of three conditions: low load ($N=27$), medium load ($N=30$), or high load ($N=28$). Participants were awarded course credit for participation.

Materials.

CRT. The original CRT was presented to participants at each test point (T1, T2, and T3). As in Study 1, a free-response format was used and scores could range from 0-3.

Matrix memory task. As in Study 1, matrix memory tasks were included as cognitive constraints. Study 2 employed low, medium, and high load conditions. Low load matrices had three coloured squares in a 3x3 grid that formed horizontal, vertical or diagonal lines. The medium load matrices had four coloured squares in 3x3 grids that formed three-piece patterns. The high-load matrices had five coloured squares in 4x4 grids (e.g., Johnson et al., 2016; Trémolière, Neys, & Bonnefon, 2012). Examples are provided in Figure 5.

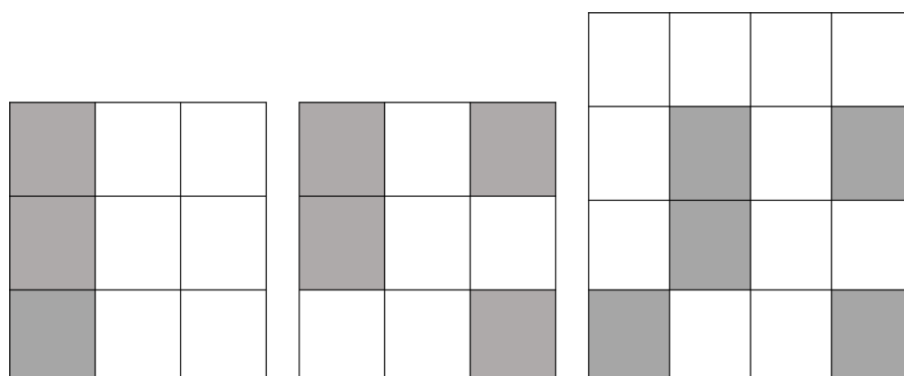


Figure 5. Examples of matrices used for (A) low, (B) medium, and (C) high load conditions.

Training items: Eighteen training items were developed to reflect the structure of the original CRT items. Training items included different numbers and content to the original CRT items to avoid participants rote-learning the correct answers. Six training

items were created per original CRT items. An example of the original CRT item and a corresponding training item is included in Table 1. Unlike the test items, training items included feedback (correct/incorrect) and guidance to help participants reach the correct solution. The feedback was tailored to the solution processes that each problem-structure required. The bat and ball item, for example, can be solved using algebra and substitution. Hence, for the six training items pertaining to the bat and ball problem, participants who gave the incorrect response were guided through a process of breaking the problem down into algebra and solving it via substitution. There are three items in the CRT, each with a different underlying mathematical structure. The problem-structure of the CRT items 2 and 3 do not have simple algebraic solutions (for example, item 3 would require an understanding of exponential growth) but they do lend themselves to worded explanations. Therefore, feedback for items reflecting problems 2 and 3 of the CRT was provided as a written explanation. See Supplementary Material for examples of feedback for each of the three problem-structures.

Numeracy. The Berlin Numeracy scale (BNS; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) was included to assess participants' pre-existing DSE. The BNS includes two to four items due to accuracy-based conditional branching. For example, if a participant gives an incorrect response they are then presented with an easier question. The BNS was developed for highly educated samples such as college students and has been used in conjunction with the CRT in several recent studies (Primi, Donati, Chiesi, & Morsanyi, 2018). An example of an item in the BNS is "Out of 1,000 people in a small town, 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir?" Scores had a possible range of 1 to 4.

WMC. The short Operation Span Task (OSPAN; Foster et al., 2015; Unsworth, Heitz, Schrock, & Engle, 2005) was used to measure WMC. The OSPAN requires

participants to remember and then recall a list of letters while intermittently assessing the validity of several short mathematical problems. A two-letter OSPAN test, for example, would start with an equation (e.g., “ $(9 / 3) - 2 = 2$?”) and the respondent must select True or False. Following this, the respondent is presented with a letter (e.g. “D”.) Next, they are presented with another equation to assess (e.g., “ $(8 / 4) - 1 = 1$?”), followed by another letter (e.g. “E”.) A recall sheet is provided at the end so that the respondent can serially recall the letters they have seen. For this example, the respondent would list D, and then E. One ‘block’ in the original OSPAN consisted of 6 trials with 2-7 letter OSPAN tests. The shorter version used in Study 2 was developed based on Foster et al.'s (2015) article, which demonstrated that performance on the full three-block OSPAN was correlated with performance on a one-block version of the task. The OSPAN task in Study 2 included 3 practice trials with letter spans of 2, and then six test trials with letter-spans of 2 to 7. Accurate serial recall was summed; scores had a possible range of 0 to 27.

Procedure. Testing was completed in person, in groups of five to eight. Participants completed the study on computers in individual partitioned booths and were provided with a notepad and pen. Once consent was obtained, participants completed a series of demographic questions. They were then given instructions for the general procedure, including instructions that the notepad could be used to assist with mathematical working out but not for the memory task components: matrix patterns or letter sequences. Notepads were checked to ensure participants did not use them for the memory tasks. Instructions were also presented with each question as appropriate, for example: “Submit your final answer only. Use numbers only (up to 2 decimal places). Exclude symbols or words e.g. \$, cents, km.” The numeracy test was completed first, then, in random order, the WMC test and the CRT training task were completed.

The CRT training task included five blocks: test 1 (T1), training 1, test 2 (T2), training 2, and test 3 (T3). Each test block included all three original CRT items. All

participants completed the test problems whilst remembering a matrix pattern. However, the complexity of the matrices was dependent on the participant's condition: low, medium, or high load. An example of the procedure of test item presentation is included in Figure 5. Different matrices were used in each test block, and for each CRT item. For each condition, the order of matrix patterns was counterbalanced, such that half of the participants received matrix patterns 1 to 9 and the other half received the same matrix patterns but in reverse order 9 to 1. This was to ensure that any training effects would not be confounded by the order of matrix patterns. The two training blocks included nine items each. None of the original CRT items were included in the training blocks. However, the nine items in each training block were created to reflect the mathematical structure of the CRT items. Each training block included three items that reflected CRT problem 1, three that reflected CRT problem 2, and three that reflected CRT problem 3. Test and training items were presented in the same order for all participants.

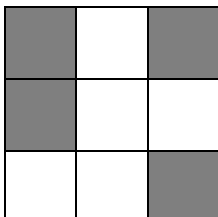
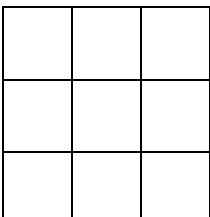
(A) “A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball.”	(B) Memorise: 	(C) “A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. <i>How much does the ball cost?</i> ”	(D) Recall: 
--	---	---	---

Figure 6. *Example of a test item from Study 2. This example is for a participant in the medium load condition at T1.*

Results

Preliminary analysis. Five participants were excluded from the analyses due to poor performance on the secondary load task across test points, meaning it could not be certain that these participants were putting genuine effort into the secondary load task. These were people with overall matrix scores more than two standard deviations below the mean score for their condition. Two from the medium load condition and three from the high load condition were excluded. Eighty participants remained in the final analyses ($N_{\text{low}}=27$, $N_{\text{medium}}=28$, $N_{\text{high}}=25$). The training paradigm was highly effective: Participants'

CRT performance increased from T1 ($M=1.10$, $SD=1.06$) to T2 ($M=2.43$, $SD=.91$), and from T2 to T3 ($M=2.57$, $SD=.79$). Participants demonstrated reasonable numeracy, ($M=2.53$, $SD=1.18$), and WMC ($M=20.46$, $SD=5.11$). Before assessing our hypotheses, we examined whether numeracy, WMC, and CRT performance at T1 differed as a function of condition. One-way between-subjects ANOVAs did not reveal any significant differences between load conditions on numeracy, $F(2, 38)=1.91$, $p=.164$, $n^2_p=.098$, WMC $F(2, 77)=.824$, $p=.442$, $n^2_p=.021$, or CRT performance at T1, $F(2, 77)=.405$, $p=.668$, $n^2_p=.010$.

Main analysis. To test our hypotheses, we used a general linear model. Results for the model are presented in Table 2. A significant three-way interaction was observed between DSE, load, and WMC on CRT performance. Therefore, while the main effects and two-way interactions are reported in Table 2, they are not examined further.

Table 2.

General Linear Model of DSE, Load and WMC on CRT Performance.

Source	Df _{Source} , Df _{Error}	F	n^2_p	p
DSE	1.32, 97.89	124.09**	.626	<.001
Load	2, 74	2.709	.068	.073
WMC	1, 74	35.73**	.326	<.001
DSE*Load	2.65, 97.89	.309	.008	.794
DSE*WMC	1.32, 97.89	.899	.012	.372
Load*WMC	2, 74	2.90	.073	.061
DSE*Load*WMC	2.65, 97.89	3.809*	.093	.016

Note. *Greenhouse-Geisser effects are report where assumptions of sphericity were violated.*

Using a generalised linear model allowed for the inclusion of WMC as a continuous predictor, rather than requiring the participants to be grouped, for example by a median split method. To examine the three-way interaction, we analysed the two-way

interactions between DSE and load on CRT performance, first, with the model adjusted to low WMC (-.5 SD), and second, with the model adjusted to high WMC (+.5 SD). These results are presented in Figure 6.

Low WMC. For low WMC, the effect of load was not significant at T1, $F(2,74)=.44$, $p=.645$, $n^2_p=.012$. However, it was significant at T2, $F(2,74)=4.83$, $p=.004$, $n^2_p=.115$, and T3, $F(2,74)=5.81$, $p=.005$, $n^2_p=.136$. Tests of simple effects revealed that at T2 the mean performance for participants in the low load condition ($M=2.53$, $SE=.16$) was greater than that of those in the medium load condition ($M=1.77$, $SE=.19$), $F(1, 74)=9.57$, $p=.003$, $n^2_p=.115$, but no different than that for the high load condition ($M=2.17$, $SE=.17$), $F(1,74)=2.54$, $p=.115$, $n^2_p=.033$. The mean performance for participants in the medium load condition was no different to that of those in the high load condition, $F(1,74)=2.46$, $p=.121$, $n^2_p=.032$. At T3 the mean performance for participants in the medium load condition ($M=1.89$, $SE=.16$) was lower than both the low load condition ($M=2.56$, $SE=.13$), $F(1, 74)=10.25$, $p=.002$, $n^2_p=.122$, and the high load condition, ($M=2.49$, $SE=.14$), $F(1, 74)=7.91$, $p=.006$, $n^2_p=.097$. There was no significant difference between the low load condition and the high load condition, $F(1,74)=.11$, $p=.739$, $n^2_p=.001$.

High WMC. In contrast to low WMC, when the model was evaluated at high WMC, the interaction between load and DSE on CRT performance was not significant at T1, $F(2,74)=1.594$, $p=.210$, $n^2_p=.041$, T2, $F(2,74)=.324$, $p=.724$, $n^2_p=.009$, or T3, $F(2,74)=1.098$, $p=.339$, $n^2_p=.029$.

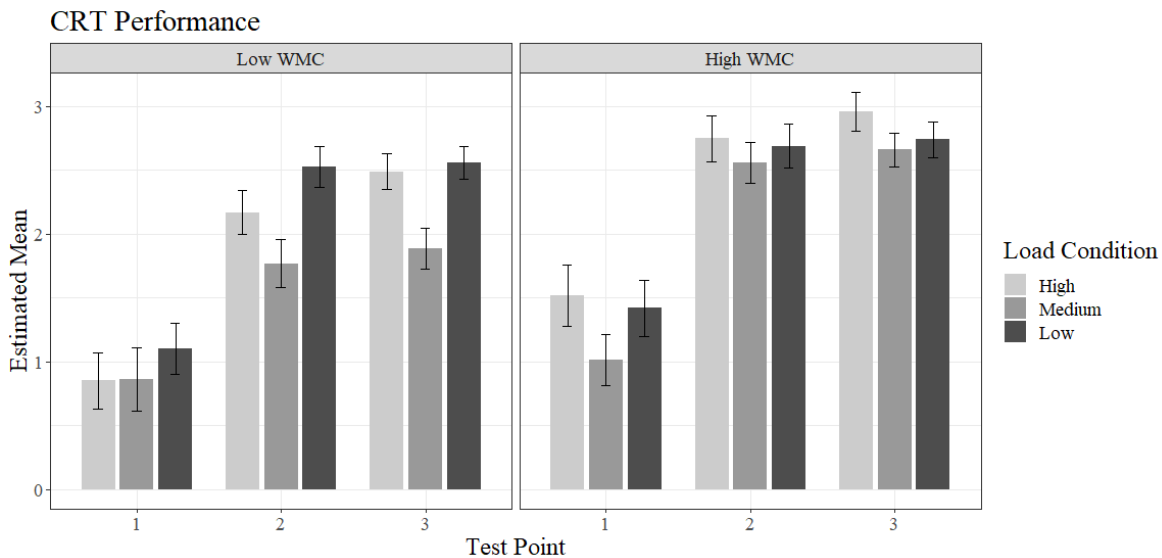


Figure 6. Means for CRT performance by load condition and test point with the model adjusted to low WMC versus high WMC. Error bars reflect $\pm 1SE$.

Discussion

Study 2 explored whether Type 2 processes can become Type 1 processes over time. This was achieved by examining the effects of DSE, WMC, and cognitive load on CRT performance. We expected that performance would increase with DSE (i.e., over training), and that DSE, load and WMC would interact to affect performance. Both hypotheses were supported. For participants with low WMC, the medium load constraint had a detrimental effect on performance at T2 and T3. This suggests that the training was effective in teaching the low WMC participants how to solve the problems but was not effective in automating the solution process. From a dual process perspective, this indicates that low WMC participants were employing Type 2 processes and did not exhibit a transition from Type 2 to Type 1 processing at T2 and T3.

In contrast, participants with high WMC were able to complete the problems with the concurrent load after just one block of training. This could be due to a dynamic effect—faster transitions from Type 2 to Type 1 stemming from the relationship between WMC and learning; or, a static effect—greater WMC reducing the impediment caused by the concurrent task; or, a combination of these factors. Given the inclusion of more

difficult tasks and the effective use of similar tasks in previous studies (Trémolière et al., 2012), it seems likely that these results are driven, at least partly, by a dynamic learning effect. Additionally, previous studies on syllogistic reasoning suggest that the effect of matrix tasks as a cognitive burden, when employed at a single point in time (and hence, point of DSE) is not mediated by working memory capacity (De Neys, 2006). Therefore, high WMC participants—like their low WMC counterparts—appear to have learned how to solve the problems but, additionally, exhibited a transition to Type 1 processing (i.e., automatisation).

From these findings branch two key implications for the use of DSE manipulations in reasoning research. First, they build on Study 1 in providing support for the suggestion that Type 2 process can transition into Type 1 processes. Second, it suggests that this method can elicit both Types of processing from the one individual and, hence, facilitate the future examination of intra-individual factors, such as conflict, in reasoning. These implications offer encouragement for the continued development and exploration of DSE-based manipulations to examine dual reasoning processes and their transition from Type 2 to Type 1.

General Discussion

In two studies, manipulations of DSE and cognitive constraints facilitated the elicitation and examination of different Types of thinking, and the transition from Type 2 to Type 1 thinking. The results demonstrated effects consistent with the claim that people change thinking Types with increases in DSE. In Study 1, we demonstrated differences in thinking Types used by participants completing the CRT across real-world changes in mathematical DSE operationalised via university degree and occupation. In Study 2, we demonstrated differences in thinking Types across changes in mathematical DSE operationalised via training. Together, these studies suggest that real-world changes in DSE are associated with differences in thinking Types and that these differences can be

generated in experimental settings.

Previous studies have employed cognitive constraints to test which Type of thinking is engaged when people are solving the CRT. Some have found support for the assertion that correct solutions to the CRT require Type 2 processing (Johnson et al., 2016) while others have found evidence that correct solutions can also be reached via Type 1 processing (Bago & De Neys, 2019). Considered alongside the present findings, we suggest that DSE (mathematical experience in particular) may, at least in part, account for these disparate findings. Those studies that showed a detrimental effect of cognitive constraint on CRT performance may have used a sample that included participants at an intermediate phase of DSE. That is, they were able to reach the solution, but the required process had not been automated. In contrast, instances when constraints did not impede performance may indicate that the sample had generally higher DSE. Therefore, they had sufficiently automated the processes such that they could perform under constraint. The current article suggests that DSE may, in part, account for previously disparate findings.

Studies that have demonstrated logical responding on bias tasks under Type 1 conditions have been presented as a challenge to the corrective interpretation of findings bias task studies (e.g., Bago & De Neys, 2017, 2018). The corrective position asserts that incorrect default responses generated via Type 1 processes can be overridden and corrected by Type 2 processes (e.g., for the CRT: Kahneman, 2011; Kahneman & Frederick, 2005). However, two-response paradigms have indicated that some participants who gave correct responses under Type 2 conditions were also able to give correct responses under Type 1 conditions (e.g., Pennycook & Thompson, 2012; Thompson & Johnson, 2014; Thompson, Prowse Turner, & Pennycook, 2011). Correct responding under Type 1 conditions suggests that such responses were not dependent on Type 2 processes; hence, it could be argued that any corrective process was either occurring through Type 1 processes or not occurring at all. However, by incorporating the parabolic relationship between DSE and thinking Type

demonstrated in the current article, a more nuanced approach can be developed for interpreting correct versus incorrect responding on bias-task findings.

The present findings suggest there may be a learning trajectory – from Type 1 to Type 2 and back to Type 1. That is, participants with low DSE may give incorrect solutions using Type 1 processes, those with intermediate DSE may be able to reach the correct solution but still require—potentially corrective—Type 2 processing, and those with high DSE may be able to complete the tasks using Type 1 processing. Studies conducted at a single point in time seem more likely to capture individuals outside the critical points of learning at which working memory plays a crucial role. An artefact of non-DSE designs is that they will have captured a sample made up largely of people with either low or high DSE, unless, for example, the participants were engaged in learning in the relevant domain. Subsequently, as reflected in previous studies, they are likely to see largely incorrect (potentially from low DSE participants) or correct (potentially from high DSE participants) Type 1 responding. Future studies should examine the potential for corrective Type 2 processing at intermediate points of DSE, for example, by combining a DSE-manipulation with the two-response paradigm.

Incorporating within-subject DSE-manipulations (i.e., with training) not only increases the likelihood of observing Type 2 processing, but also presents a promising method for the closer examination of intra-individual factors that may be involved in engaging Type 2. Metacognitive factors like *conflict* and *feeling of rightness* are increasingly featured in theoretical and empirical work (e.g., De Neys, 2012; Thompson et al., 2011). The logical intuition model, for example, proposes that the engagement of Type 2 processing may be related to conflict generated by the relative activation of two or more competing Type 1 processes (De Neys, 2012, 2014; Bago & De Neys, 2017). The metacognitive model proposes that the engagement of Type 2 processing is associated with a lowered feeling of rightness stemming from the difficulty with which an initial Type 1

process is activated and brought to mind (Thompson et al., 2011). Studies that explore the relationship between intra-individual factors and Type of thinking that are conducted at a single point in time—and, hence a single point on the DSE trajectory—are likely to observe participants consistently employing one Type of processing or the other. By implementing training paradigms, the chances of observing individual participants changing between Types of processing, in other words, studies that have instance of participants using both Type 1 and Type 2 thinking, is increased. Therefore, expanding the potential for examining modern dual process models that include intra-individual factors.

This article presents a unique combination of DSE-manipulations and cognitive constraints to explore the relationship between these factors and performance on the ubiquitous CRT. The studies incorporated ideas from the expertise and working memory domains to produce findings that, as far as we know, are the first to provide explicit empirical support for the dual process assumption that Type 2 processes can become Type 1 processes with practice. Additionally, the findings suggest that DSE may be an important factor in accounting for previously inconsistent findings. By harnessing a common human experience—transitions from effortful to easy processing—this article introduces theoretical and methodological advances to dual process reasoning research.

References

- Allport, D. A., Antonis, B., & Reynolds, P. (1972). On the Division of Attention: A Disproof of the Single Channel Hypothesis. *Quarterly Journal of Experimental Psychology*, 24(2), 225–235. <https://doi.org/10.1080/00335557243000102>
- Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.
- Baddeley, A. (1986). *Working memory*. Clarendon Press.
- 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*, 1–43. <https://doi.org/10.1080/13546783.2018.1507949>
- Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, 14(1), 12–23. <https://doi.org/10.1037/0096-1523.14.1.12>
- Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*, 50(5), 1953–1959. <https://doi.org/10.3758/s13428-017-0963-x>
- Buckner, R. L. (2004). Memory and Executive Function in Aging and AD. *Neuron*, 44(1), 195–208. <https://doi.org/10.1016/j.neuron.2004.09.006>
- Chase, W. G., & Ericsson, K. A. (1982). Skill and Working Memory (pp. 1–58). [https://doi.org/10.1016/S0079-7421\(08\)60546-0](https://doi.org/10.1016/S0079-7421(08)60546-0)
- Cokely, E., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making*, 4(1), 20–33.

- Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment & Decision Making*, 7(1), 25–47. Retrieved from <http://simsrad.net.ocs.mq.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=71325917&site=ehost-live>
- De Neys, W. (2006). Dual Processing in Reasoning. *Psychological Science*, 17(5), 428–433. <https://doi.org/10.1111/j.1467-9280.2006.01723.x>
- De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. Retrieved from <http://pps.sagepub.com>
- 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. (2017). *Dual process theory 2.0. Dual Process Theory 2.0*. Routledge. <https://doi.org/10.4324/9781315204550>
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, 128(3), 309–331. <https://doi.org/10.1037/0096-3445.128.3.309>
- Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, 49(8), 709–724. <https://doi.org/10.1037/0003-066X.49.8.709>
- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2), 211–245. <https://doi.org/10.1037/0033-295X.102.2.211>
- 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>
- Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W.

- (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, 43(2), 226–236. <https://doi.org/10.3758/s13421-014-0461-7>
- Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Gigerenzer, G. (2007). Gut Feeling: The Intelligence of The Unconscious. *New Penguin Books*, 280.
- Glanzer, M., Dorfman, D., & Kaplan, B. (1981). Short-term storage in the processing of text. *Journal of Verbal Learning and Verbal Behavior*, 20(6), 656–670. [https://doi.org/https://doi.org/10.1016/S0022-5371\(81\)90229-2](https://doi.org/https://doi.org/10.1016/S0022-5371(81)90229-2)
- Glanzer, M., Fischer, B., & Dorfman, D. (1984). Short-term storage in reading. *Journal of Verbal Learning and Verbal Behavior*, 23(4), 467–486. [https://doi.org/https://doi.org/10.1016/S0022-5371\(84\)90300-1](https://doi.org/https://doi.org/10.1016/S0022-5371(84)90300-1)
- Gobet, F., & Simon, H. A. (1996). Templates in Chess Memory: A Mechanism for Recalling Several Boards. *Cognitive Psychology*, 31(1), 1–40. <https://doi.org/10.1006/cogp.1996.0011>
- Guida, A., Gobet, F., Tardieu, H., & Nicolas, S. (2012). How chunks, long-term working memory and templates offer a cognitive explanation for neuroimaging data on expertise acquisition: A two-stage framework. *Brain and Cognition*, 79(3), 221–244. <https://doi.org/10.1016/j.bandc.2012.01.010>
- Hambrick, D. Z., & Engle, R. W. (2002). Effects of Domain Knowledge, Working Memory Capacity, and Age on Cognitive Performance: An Investigation of the Knowledge-Is-Power Hypothesis. *Cognitive Psychology*, 44(4), 339–387. <https://doi.org/10.1006/cogp.2001.0769>
- Hambrick, D. Z., & Engle, R. W. (2003). The role of working memory in problem solving. *The Psychology of Problem Solving*, 176–206.

- Hambrick, D. Z., & Oswald, F. L. (2005). Does domain knowledge moderate involvement of working memory capacity in higher-level cognition? A test of three models. *Journal of Memory and Language*, 52(3), 377–397.
<https://doi.org/10.1016/j.jml.2005.01.004>
- Handley, S. J., & Trippas, D. (2015). Dual Processes and the Interplay between Knowledge and Structure: A New Parallel Processing Model. *The Psychology of Learning and Motivation*. San Diego : <https://doi.org/10.1016/bs.plm.2014.09.002>
- Hebb, D. O. (Donald O. (1949). *The organization of behavior : a neuropsychological theory / D.O. Hebb*. New York: Wiley.
- Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, 87(1), 18–32. <https://doi.org/10.1037/0022-0663.87.1.18>
- Johnson, E. D., Tubau, E., & De Neys, W. (2014). The unbearable burden of executive load on cognitive reflection: A validation of dual process theory The unbearable burden of executive load on cognitive reflection: A validation of dual process theory. *Proceedings of the Annual Meeting of the Cognitive Science Society*, (36), 36.
- Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, 164, 56–64.
<https://doi.org/10.1016/j.actpsy.2015.12.008>
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: 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: Cambridge University Press.
- Keren, G. (2013). A Tale of Two Systems. *Perspectives on Psychological Science*, 8(3), 257–262. <https://doi.org/10.1177/1745691613483474>
- Klein, G. (2003). *The Power of Intuition: How to Use Your Gut Feelings to Make Better*

- Decisions at Work*. Crown Business.
- Kruglanski, A. W. (2013). Only One? The Default Interventionist Perspective as a Unimodel—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, 8(3), 242–247. <https://doi.org/10.1177/1745691613483477>
- Kyllonen, P. C., & Dennis, A. (1996). Is working memory capacity Spearman's *g*. *Human Abilities: Their Nature and Measurement*, 49–75.
- Kyllonen, P. C., & Stephens, D. (1990). Cognitive abilities as determinants of success in acquiring logic skill. *Learning and Individual Differences*, 2(2), 129–160. [https://doi.org/10.1016/1041-6080\(90\)90020-H](https://doi.org/10.1016/1041-6080(90)90020-H)
- Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual Differences in Numeracy and Cognitive Reflection, with Implications for Biases and Fallacies in Probability Judgment. *Journal of Behavioral Decision Making*, 25(4), 361–381. <https://doi.org/10.1002/bdm.752>
- Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, 130(4), 621–640. <https://doi.org/10.1037/0096-3445.130.4.621>
- Obrecht, N. A., Chapman, G. B., & Gelman, R. (2009). An encounter frequency account of how experience affects likelihood estimation. *Memory & Cognition*, 37(5), 632–643. <https://doi.org/10.3758/MC.37.5.632>
- Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, 11(6), 988–1010. <https://doi.org/10.3758/BF03196730>
- Osman, M. (2013). A Case Study: Dual-Process Theories of Higher Cognition—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, 8(3), 248–252. <https://doi.org/10.1177/1745691613483475>
- Park, D. C., Smith, A. D., Lautenschlager, G., Earles, J. L., Frieske, D., Zwahr, M., &

- Gaines, C. L. (1996). Mediators of long-term memory performance across the life span. *Psychology and Aging, 11*(4), 621–637. <https://doi.org/10.1037/0882-7974.11.4.621>
- 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>
- Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review, 19*(3), 528–534. <https://doi.org/10.3758/s13423-012-0249-3>
- Primi, C., Donati, M. A., Chiesi, F., & Morsanyi, K. (2018). Are there gender differences in cognitive reflection? Invariance and differences related to mathematics. *Thinking & Reasoning, 24*(2), 258–279. <https://doi.org/10.1080/13546783.2017.1387606>
- Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory. *Judgment & Decision Making, 7*(3), 1–45.
- Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology, 6*, 532. <https://doi.org/10.3389/fpsyg.2015.00532>
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22. <https://doi.org/10.1037/0033-2909.119.1.3>
- Stuppelle, E., Gale, M., & Richmond, C. (2013). Working memory, cognitive miserliness and logic as predictors of performance on the cognitive reflection test. *Proceedings of the ... Annual Conference of the Cognitive Science Society., 35*(35), 1396–1401.
- Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning, 20*(2), 215–244. <https://doi.org/10.1080/13546783.2013.869763>
- 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>
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39(7), 1275–1289. <https://doi.org/10.3758/s13421-011-0104-1>
- Trémolière, B., Neys, W. De, & Bonnefon, J.-F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition*, 124(3), 379–384. <https://doi.org/10.1016/j.cognition.2012.05.011>
- Trippas, D., & Handley, S. J. (2017). The parallel processing model of belief bias: Review and extensions. In W. De Neys (Ed.), *Dual process theory 2.0* (1st ed.). London: Routledge. <https://doi.org/10.4324/9781315204550>
- Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37(3), 498–505. <https://doi.org/10.3758/BF03192720>
- Van Lier, J., Revlin, R., & De Neys, W. (2013). Detecting Cheaters without Thinking: Testing the Automaticity of the Cheater Detection Module. *PLoS ONE*, 8(1), e53827. <https://doi.org/10.1371/journal.pone.0053827>
- Verschueren, N., Schaeken, W., & d’Ydewall, G. (2004). Everyday Conditional Reasoning with Working Memory Preload. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 26(26), 1399–1404. Retrieved from <https://cloudfront.escholarship.org/dist/prd/content/qt7kk1x3qx/qt7kk1x3qx.pdf>
- Wastell, C. A. (2014). An emergence solution to the reasoning dual processes interaction problem. *Theory & Psychology*, 24(3), 339–358.

<https://doi.org/10.1177/0959354314533442>

Welsh, M. B., Burns, N. R., & Delfabbro, P. H. (2013). The Cognitive Reflection Test: how much more than Numerical Ability? *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35(35), 1587–1592. Retrieved from <https://cloudfront.escholarship.org/dist/prd/content/qt68n012fh/qt68n012fh.pdf>

Paper 2

No Pain, No Gain: Does Cognitive Conflict Predict Learning and Effortful Reasoning?

Zoe A. Purcell¹, Colin A. Wastell¹, Naomi Sweller¹, and Stephanie Howarth²

¹Department of Psychology, Macquarie University, Sydney

²Department of Cognitive Science, Macquarie University, Sydney

This paper has been submitted for publication to
*The Journal of Experimental Psychology: Learning, Memory, and
Cognition*

Abstract

The dual process logical intuition model asserts that metacognitive conflict is associated with the engagement of effortful thinking (De Neys, 2012, 2014). As a learner gains more Domain Specific Experience (DSE), their thinking style transitions between intuitive and effortful (working memory dependent) processing (Purcell, Wastell, & Sweller, n.d.³). In two studies, we test whether the changes in thinking, elicited via a DSE training paradigm, are associated with shifts in cognitive conflict for participants solving the cognitive reflection test (CRT; Frederick, 2005). In Study 1 we employ a concurrent load task; participants are required to solve the CRT while completing a memory task of varying complexity. In Study 2 we employ a two-response paradigm; participants are required to solve each problem twice, first, under a short deadline, and second, without a timing restriction. The effects of these cognitive constraints are examined alongside confidence-based conflict, changes in DSE, and performance on the CRT. The main findings indicate that conflict is associated with learning and predicts effortful—working memory dependent—thinking.

³ Paper 1 in the current thesis.

No Pain, No Gain: Does Cognitive Conflict Predict Learning and Effortful Reasoning?

“Reasoning” carries connotations of an effortful process of deep consideration and contemplation. One might imagine a process of learning something new or weighing-up pros and cons. These processes entail some form of ambiguity, for example, if we have positives and negatives to weigh-up in order to reach a decision we may be forced to “sit on the fence” and consider our options before proceeding; or, in the case of learning, we are confronted with a situation we don’t yet know how to deal with, or a problem that we don’t yet know how to solve. In these instances, it would not be surprising if the reasoner experienced a sense of uncertainty and (uncomfortable) awareness of the fact that something is unknown. That sense of uncertainty, some reasoning scholars have proposed, may not just be associated with these instances of deeper thought but could in fact play a part in the engagement of effortful thinking.

Dual process theories suggest that there are two distinct Types of thinking (e.g., (Evans & Stanovich, 2013; Kahneman, 2011; Sloman, 1996); Type 1, which is fast, automatic, and does not require working memory (for example, when an adult attempts the problem “ $2+2=?$ ”), and Type 2 which is slower, deliberative, and requires working memory (for example, when an adult attempts the problem “ $363+58=?$ ”). Of the many dual process theories, we focus on the logical intuition dual process model because of the emphasis it places on the role of uncertainty—caused by cognitive conflict—and its potential for triggering Type 2 processes (De Neys, 2012, 2014). The logical intuition model asserts that multiple Type 1 processes can be simultaneously engaged in response to a particular situation or problem (see also Pennycook, Fugelsang, & Koehler, 2015). When two or more Type 1 processes are triggered, their relative activation can produce a sense of conflict which leads to Type 2 engagement (De Neys, 2012, 2014). This model is depicted in Figure 1.

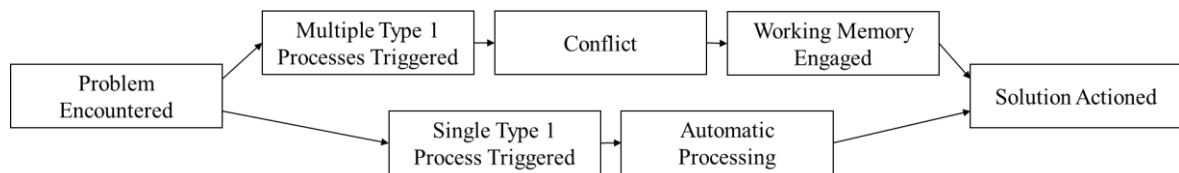


Figure 1. *Conjectural model of the logical intuition dual process theory.*

This idea is commonly illustrated using the bat-and-ball problem. The bat-and-ball problem is one of three items in the CRT (Frederick, 2005); it states: “Together, a bat and a ball cost \$1.10. The bat costs \$1 more than the ball. How much does the ball cost?” The answer that springs to mind for many respondents is 10c (Kahneman, 2011). However, the correct answer is 5c. If the ball cost 5c, the bat would cost \$1.05, their summed value would therefore amount to \$1.10. Under the logical intuition model, the bat-and-ball problem is thought to be able to activate two competing Type 1 processes, one leading to the heuristic “10c” response, and the other leading to the logical “5c” response (De Neys, Rossi, & Houdé, 2013). If two Type 1 processes are activated, the logical response could play a role in generating cognitive conflict (i.e., a sense of uncertainty), even if the heuristic (incorrect) response is ultimately the one expressed. That is, even though the heuristic process had the greatest activation, the logical process may have been sufficiently activated to compete with the heuristic and produce a sense of conflict. That conflict may lead to the engagement of Type 2 thinking and, in some cases, allow the reasoner to engage their working memory to effortfully override the incorrect response. This proposed relationship is examined in the empirical components of this article which combine methods from conflict detection and cognitive constraint studies.

Conflict detection studies examine the relationship between people’s reported experience of cognitive conflict and their performance on bias tasks. These studies often include “conflict items” which have two plausible cognitive outputs (e.g., the bat-and-ball problem mentioned earlier which has two common outputs: 5c and 10c). These conflict items are contrasted with “no-conflict items” which have only one plausible outcome. A

no-conflict version of the bat-and-ball problem, for example, may state: “Together, a bat and a ball cost \$1.10. The bat costs \$1. How much does the ball cost?” The only plausible answer for this problem is 10c. By contrasting conflict and no-conflict items, studies have demonstrated that even when respondents answer with erroneous heuristic responses, like 10c, they show signs of conflict sensitivity. For example, some studies have found longer response times on conflict items (e.g., Bonner & Newell, 2010; De Neys & Glumicic, 2008; Pennycook, Fugelsang, & Koehler, 2012; Stupple, Ball, & Ellis, 2013) and others have found that participants are less confident on the conflict items than the no-conflict versions (e.g., De Neys, Cromheeke, & Osman, 2011; De Neys et al., 2013; De Neys & Feremans, 2013; Gangemi, Bourgeois-Gironde, & Mancini, 2015; Johnson, Tubau, & De Neys, 2016; Thompson & Johnson, 2014). These results have been interpreted as indications that people can implicitly detect the conflict between competing processes (Bago & De Neys, 2019; De Neys, 2012).

Previous studies have examined whether conflict is present for reasoners solving the bat-and-ball problem. One measure used to examine conflict is cursor trajectory. Travers, Rolison and Feeney (2016) employed a mouse-tracking paradigm with a multiple-choice version of the bat-and-ball problem to determine if participants were tempted by the heuristic/logical options. The cursor movements indicated that, even when participants gave the correct response, they were often tempted by the heuristic option. However, those giving heuristic, incorrect responses were not attracted by the logical option. This suggests that correct but not incorrect responding may be accompanied by conflict and deliberation.

Bago, Raelison and De Neys (2019) also examined conflict during reasoning on the CRT but with a different approach. Participants were first asked to solve the bat-and-ball like problem by generating their own response (i.e., free-response format). Next, they were asked to attempt the problem again, however, this time they were only offered non-heuristic options in a multiple-choice format (e.g., 1c, 5c, and 15c). At first attempt,

participants often gave an incorrect response (e.g., 10c), however, on their second attempt (when the heuristic option was not available) they were more likely to guess that the correct response was smaller than their previous, incorrect response (e.g., 1c or 5c). Although this pattern of results might be explained by alternative phenomena, they are consistent with the claim that the participants may have some intuition regarding the mathematical principle behind the problem and that they may have some underlying conflict between the process leading to the correct response and another leading to the incorrect response.

Conflict studies lend considerable support to the logical intuition model's claim that multiple processes can be triggered in Type 1 processing and that these processes may compete and produce a sense of cognitive conflict. However, it is less clear whether the conflict produced by competing Type 1 processes is involved in engaging Type 2 thinking. To investigate whether conflict is associated with Type 2 engagement, some studies have included both measures of conflict and manipulations of "cognitive constraints". Manipulating cognitive constraints is thought to impact participants' use of Type 2 processes. Type 2 processes are cognitively demanding and require executive resources like working memory (e.g., Evans & Stanovich, 2013; Kahneman, 2011). Cognitive constraints usually involve imposing an additional memory load (simultaneous to the reasoning task), a timing deadline, or both, to impede Type 2 processing while a participant is completing a reasoning task (e.g., Bago & De Neys, 2017, 2019; Johnson et al., 2016). These paradigms work on the assumption that when the resources necessary for Type 2 processing are compromised, the reasoner should rely on the less-demanding Type 1 processes.

The two-response paradigm, developed by Thompson, Turner and Pennycook (2011), is a well-known cognitive constraint method. First, participants are instructed to respond to a problem as quickly as possible, sometimes under timed conditions, and

second, participants respond to the same problem but can take as long as they need. It is expected that, during the first attempt, participants do not have time to engage working memory-dependent Type 2 processes and so are more likely to use Type 1 processes and arrive at the heuristic response. During the second attempt, participants are at leisure to engage working memory-dependent, deliberative Type 2 processes should be more likely to arrive at the correct response. However, Thompson and colleagues have found that when a participant is able to successfully complete a problem on the second attempt (i.e., when encouraged to take as much time as needed), they can generally provide the correct response at the first attempt as well (e.g., Bago & De Neys, 2019; Pennycook & Thompson, 2012; Thompson & Johnson, 2014; Thompson et al., 2011). That is, many reasoners who reach the correct solution manage to do so without deliberation (Bago & De Neys, 2019). The lack of answer change between the first and second attempts has been taken as indication that the reasoners producing correct responses were able to do so using Type 1 processes, without engaging their working memory, and hence, it is possible that they possess intuitive logical processes (Bago & De Neys, 2019; Pennycook & Thompson, 2012; Thompson & Johnson, 2014; see also Bago & De Neys, 2017; Newman, Gibb, & Thompson, 2017).

Two-response paradigms have been used to examine the relationship between Type 2 thinking and conflict. Bago and De Neys (2019) examined whether cognitive constraints impact performance on the bat-and-ball problem by employing a variety of two-response paradigms. They found that participants were unlikely to change their responses between first and second attempt; those who did change their response reported lower confidence on conflict relative to no-conflict items. That is, answer change was associated with higher conflict. However, answer changes were rare and occurred on only 7.42% of trials. Furthermore, Bago and De Neys (2019) reported high stability indices. That is, that 91.3% of participants remained in the same change category across trials; the participants were

either likely to change responses on every trial, or to not change responses on any trial.

Even when subjects have been repeatedly presented with the bat and ball problem during a two-response paradigm, very few participants have exhibited answer changes. Raoelison and De Neys (2019) presented participants with 50 versions of the bat-and-ball problem, with the same problem structure but different content and quantities. The majority of participants continued to give the incorrect answer at both first and second response, throughout the study (61%). However, a small portion of participants showed evidence of learning—they switched from dual incorrect answers to dual correct answers. Of this portion of participants, most recorded just one trial on which they exhibited deliberation (i.e. a change from an incorrect answer at first response to a correct answer at second response). This indicates that for some participants the bat and ball problem might be classified as an insight task that invoked “spontaneous learning”; for most, however, the heuristic response remained persistent and difficult to override without explicit instruction.

Limited cases of deliberation and high stability indices where most participants consistently exhibit only one of the two Types of thinking can be a concern for researchers wanting to examine within-individual factors associated with answer change (see also Bago & De Neys, 2017). High stability indices can naturally generate between-subject comparisons where participants are either exhibiting mostly Type 1 or mostly Type 2 thinking in spite of within-subject designs. When examining the factors that are associated with Type 1 versus Type 2 thinking, it is helpful if the individual participants exhibit both patterns of thinking. Overall, the conflict and cognitive constraint studies lend support to the logical intuition model and its assertions about conflict and Type 2 thinking. However, that support is limited because instances of Type 2 engagement (indicated by answer changes) seem relatively uncommon and participants tend to exhibit high stability indices.

A recent article by Purcell et al. (n.d.⁴) explored the relationship between domain-specific experience (DSE), working memory engagement, and performance on the CRT. They demonstrated that as experience increased, dependency on working memory was higher at intermediate levels of experiences (i.e., during learning phases) and then decreased with continued practice. This paradigm may, therefore, assist in reducing stability indices and increasing instances of Type 2 thinking. The following two studies employ this paradigm to examine whether experiences of conflict are related to (training induced) changes in thinking Type. We aim to investigate whether the common experience of cognitive unease during learning might be reflected in a relationship between conflict, Type 2 thinking, and performance on the CRT.

In both studies we employed a training paradigm to increase DSE with four test points (T1, T2, T3 and T4) and three training blocks (between the test points). At each test point cognitive constraints were imposed. In Study 1 the constraint involved a simultaneously imposed memory task, and in Study 2 the constraint involved a timing deadline. In Study 1, at each test point, a matrix memory task was imposed wherein participants completed the CRT problems with either a low or high memory load. We expected that the higher the load, the poorer the performance. In line with Purcell et al. (n.d.⁵), we expected that subsequent to the first block of training, as DSE increased, performance would be less impeded by the loads due to increasing automaticity. That is, the constraint was expected to have a greater impact at T2 and T3 than T1 (before training) and T4 (after automation). We also expected participants who self-reported higher cognitive conflict at a particular test point would also experience decreased performance when under higher memory load. That is, in line with the logical intuition model, conflict was expected to be positively associated with working memory engagement.

⁴ Paper 1 in the current thesis.

⁵ Paper 1 in the current thesis.

In Study 2, we employed a two-response paradigm wherein participants were required to respond to each CRT problem twice. At first response, under a time constraint, and at second response, with no time constraint. We expected that participants would be more likely to exhibit response changes at T2 and T3 than T1 (before training) and T4 (after automation). We also expected that higher conflict scores would predict a greater chance of answer change (from Response 1 to Response 2). That is, as for Study 1, conflict was expected to be associated with working memory engagement.

Study 1

Study 1 examined the effects of constraint, DSE, and conflict on CRT performance. The study employed a matrix memory load to manipulate cognitive constraint in conjunction with a mathematical training paradigm to manipulate DSE. Participants' confidence in their answers was used to generate a measure of conflict. Additional measures were taken to control for individual differences and check for potential confounds. Participants' numeracy was expected to reflect their mathematical DSE and, subsequently, influence their point of automation; the higher the numeracy the earlier their expected point of automation; therefore, numeracy was included as a covariate. Participants' working memory capacity was expected to influence the effect of the matrix memory loads; the higher a participant's working memory capacity, the lower the expected effect of load. Additionally, intelligence and mathematical anxiety were measured to ensure these factors did not confound with the between-subjects load manipulation. Mathematical anxiety and intelligence have been linked to CRT performance (e.g., Primi, Donati, Chiesi, & Morsanyi, 2018; Toplak, West, & Stanovich, 2011). Based on Purcell et al.'s (n.d.⁶) study, we expected that DSE, constraint, and working memory capacity would interact to affect performance on the CRT. In line with conflict studies (e.g., Bago & De

⁶ Paper 1 in the current thesis.

Neys, 2019), we expected that conflict, constraint, and DSE would interact to predict performance on the CRT. That is, we expected that after the initial test point as DSE increased, the effect of constraint would decrease (reflecting a process of automation), and this shift would be moderated by decreasing conflict scores.

Method

Participants and Design. We used a 4 (test point) x 2 (low load or high load) mixed design. Participants were recruited via Amazon's Mechanical Turk platform and redirected to the study on the Qualtrics platform. Mechanical Turk's exclusion capabilities were used to sample participants with a 95% or better performance record on previous Mechanical Turk studies. The Qualtrics exclusion capabilities were used to prevent participants participating more than once based on IP address. Participants were 107 Mechanical Turk workers, paid \$US6 for participating, from the United States (78.5%) and India (21.5%). Each participant was randomly allocated to low load ($N=61$) or high load ($N=46$) constraint conditions. The sample was comprised of 46.7% female, 52.3% male and .9% unspecified/other, with ages ranging from 22 to 66 ($M=39.44$, $SD=11.00$). Participants were required to have a minimum of high school level mathematics to be eligible for the study. Of these participants, 48% had completed a Bachelor's/Associate/Postgraduate Degree, and 14% had partially completed a Bachelor's Degree or similar.

Materials. Maths Problems. The studies contained conflict and no-conflict items. Conflict items were developed to reflect the structures of the original 3-item CRT (Frederick, 2005). An example of a conflict test item is: "A bag and a badge cost \$12.10 in total. The bag costs \$2.00 more than the badge. How much does the badge cost?". No-conflict items were similar to the conflict items, but the process that would have resulted in the incorrect intuitive response on the conflict item (e.g., $12.10 - 2.00 = 10.10$) would, for no conflict items, result in correct responses. An example of a no conflict test item is: "A

magazine and a banana together cost \$2.90. The magazine costs \$2. How much does the banana cost?”. Study 1 employed a free-response format; participants entered numbers only, no options were provided, and no time limit was imposed.

There were seven blocks of maths problems: four blocks of test items and three blocks of training items. The order of blocks was 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). Cognitive constraints (matrices) were imposed on test but not training items. Training items included feedback (correct/incorrect) and guidance (see Supplementary Material). To allow for the fact that although they had the same structure, some items may be more difficult than others, the order of the blocks was counterbalanced such that the items used in T1 for one half of the participants were the same items used in T4 for the other half of the participants and the items used for T2 for one half of the participants were the same items used in T3 for the other half of the participants. Therefore, shifts in performance over time were less likely to reflect systematic differences in the difficulty of the items. The order of the items was also randomised within each block. The no-conflict items were included to create conflict scores (see below), however, only the conflict items were used for calculating the dependent variables. Performance scores ranged from 0-3 for each of the four test points.

Cognitive Constraint. Matrices were used as cognitive constraints. Participants were required to memorise 3x3 matrices with three (low load) or four (high load) coloured squares presented for 900ms (see Figure 2). Low load matrices included “one-piece” patterns with three coloured squares in a 3x3 grid; the squares were in horizontal, vertical or diagonal lines (Bethell-Fox & Shepard, 1988). High load matrices included “three-piece” patterns with the four coloured squares. Two of squares could not be adjacent to the other two squares but could meet at the diagonal (see Figure 2). Similar matrices have been used to hinder executive resources (e.g., Bago & De Neys, 2019; Johnson, Tubau, & De

Neys, 2014). Scores on matrix tasks were calculated by adding the number of squares correctly selected. Scores could range from 0-3 for low load participants and 0-4 for high load participants.

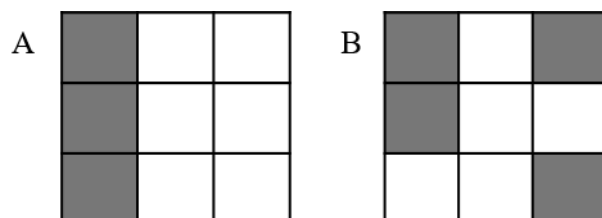


Figure 2. *Example of matrices for (A) low load and (B) high load constraint conditions.*

Conflict Factor. A conflict factor was computed for each conflict item. Previous studies have measured conflict by comparing post-response confidence on conflict and no-conflict items (e.g. Bago & De Neys, 2017; Johnson et al., 2016). Conflict items are assumed to trigger two competing responses: one leading to the correct logical solution and the other leading to an incorrect biased solution. These competing responses are thought to reduce confidence relative to no-conflict items for which there is no tempting incorrect response. After responding to each maths problem (both conflict and no-conflict), participants recorded their confidence in that response on a scale of 0 to 100 (0=not at all confident, 100=absolutely confident). The confidence rating for the conflict item was subtracted from the confidence rating on the corresponding no-conflict item, to calculate a conflict factor. Higher scores reflected greater conflict on the conflict versus no-conflict items.

Numeracy. Participants' numeracy was measured with the Berlin Numeracy Test (BNT; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). The BNT includes two to four questions. The item presentation is dependent on each participant's accuracy. For example, if a participant gives a correct response they are then presented with a harder problem. An example of an item in the BNT is "Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500

inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir?" Scores on the BNT could range from 1 to 4.

Intelligence. Participants' intelligence was measured using a 12-item version of the Advanced Raven's Progressive Matrices (RPM; Bors & Stokes, 1998). This 12-item version was developed for educated samples and is highly correlated with the full RPM ($r=.92, p<.001$; Bors & Stokes, 1998). Participants were presented with a pattern to complete and eight shapes to select from. They were asked: "Which shape completes the pattern above? Type the number corresponding to the shape you wish to select." Scores on the RPM could range from 0 to 12.

Mathematical anxiety. Mathematical anxiety was measured with the Abbreviated Mathematical Anxiety Scale (AMAS; Hopko, Mahadevan, Bare, & Hunt, 2003). The AMAS has shown good internal reliability (Cronbach's alpha of .90; Hopko et al., 2003; and .89, Primi et al., 2018), and ecological validity (Primi et al., 2018). Participants respond on a five-point Likert scale ranging from 1 = Low Anxiety to 5 = High Anxiety to phrases such as: "Having to use the tables in the back of a maths book." Scores on the AMAS could range from 9 to 45. Higher scores reflected greater mathematical anxiety. An attention-check item was embedded within the AMAS between items 7 and 8; it read "If you are paying attention, please select '5' High anxiety. Participants who did not select '5' were excluded from the final analysis.

Working Memory Capacity. Working memory capacity was measured using a short version of the Operation Span Task (OSPAN; Foster et al., 2015; Unsworth, Heitz, Schrock, & Engle, 2005). The OSPAN requires participants to verify several short mathematical equations presented in between letters they are later required to recall. For example, a two-letter OSPAN test would start with an equation (e.g., " $(9 / 3) - 2 = 2$?") to which the respondent must select True or False. Following this, the respondent is presented with a letter (e.g. "D"). They are then presented with another equation (e.g., " $(8 / 4) - 1 = 1$

?) to assess and another letter (e.g. “E”). A recall sheet is then presented. Participants recall the letters in the order they were presented (in this example: D, E). Three practice trials with two-letter spans were presented and then six test trials with letter-spans of three to eight. Scores were computed by summing the number of letters correctly recalled in the correct order (Foster et al., 2015; Turner & Engle, 1989). Scores had a possible range from 0 to 33.

Procedure. Participation was completed online via Qualtrics. After consent was obtained, participants completed a series of demographic questions. They were then given instructions for the general procedure. Instructions were also presented on each page as appropriate, for example: “Submit your final answer only. Use numbers only (up to 2 decimal places). Exclude symbols or words e.g. \$, cents, km.” Participants first completed the scales measuring mathematical anxiety, intelligence and numeracy. The order of scale presentation was randomised except for the mathematical anxiety measure, which was always presented first to prevent the numeracy or other questionnaires from impacting the measure. The order of the working memory capacity test and the test-and-training section was counterbalanced. The order of the items within each scale was randomised except for the numeracy scale which required a response-dependent presentation order (see Materials).

The test-and-training section contained three practice questions and then the seven test and training blocks (see Materials). First, participants were presented with instructions: “On the next screen you will see a grid pattern to memorize. For every grid pattern, it is very important that you remember the pattern **WITHOUT** using any external aids like a notepad.” On a new screen, participants saw a matrix pattern (see Figure 2) with the instructions: “Memorise which cells are highlighted. You will be asked to recall them soon. You **CANNOT** use a memory aid to help you remember the grid.” This was displayed for 900ms. On the following screen they were presented with the maths problem

to which they responded with a numerical answer and pressed “Next” to proceed. They were then presented with the question: “How confident are you in your answer? Please type a number from 0 (absolutely not confident) to 100 (absolutely confident).” They clicked Next to proceed. On a new screen, they were presented with a blank grid and the instructions: “Click the cells that were highlighted in the grid you saw earlier.” If a cell was clicked on, it would turn from white to green, if it was clicked a second time, it would return to white. The participant could proceed by clicking ‘Next’.

Results

Exclusion and Descriptive Analyses. Strict exclusion criteria were employed. First, participants who failed the attention check items were excluded (N=107 remaining). Second, to ensure that the cognitive constraint was properly completed, participants who scored less than 2 standard deviations below the mean for matrix task performance (calculated separately for low and high load conditions) or two standard deviations below the mean for the working memory capacity test were excluded from analysis. Consequently, 100 of the 107 participants remained in the dataset for analysis. Participants’ scores on numeracy, mathematical anxiety, intelligence and working memory capacity are reported in Table 1. To check for potential confounds we tested whether these factors differed as a function of constraint condition. There were no significant differences between conditions on numeracy, $F(1,98)=.65, p=.424, n^2_p=.007$, mathematical anxiety, $F(1,98)=.36, p=.552, n^2_p=.004$, intelligence, $F(1,98)=.074, p=.786, n^2_p=.001$, or working memory capacity, $F(1,98)=.004, p=.947, n^2_p<.001$. Averaged across conditions and test points, participants scored 7.91 (SD = 3.59) out of 12 conflict items. More detail on the participants’ mean performance on conflict items at each test point and the correlations between each of the individual difference factors are provided in the Supplementary Material (Table 2).

DSE, Constraint, WMC and Performance. A mixed ANCOVA was used to

examine the relationship between DSE (test point), constraint (load) and performance on the conflict items. The independent factors were DSE (test point, within-subjects), constraint (low or high load, between-subjects), and WMC (continuous); and a covariate, numeracy. Where sphericity was violated, Greenhouse-Geisser adjusted results are reported. The main effect of DSE on performance—when averaged across the remaining factors—was significant, $F(2.59, 245.80) = 22.83$, $p < .001$, $n^2_p = .194$. Follow-up pairwise comparisons compared each time point with the successive time point. P -values were compared to a Bonferroni-adjusted alpha of .017 to account for the three comparisons. The pairwise comparisons revealed that performance improved from T1 ($M = 1.55$, $SD = 1.11$) to T2 ($M = 2.19$, $SD = 0.93$), $F(1, 93) = 53.10$, $p < .001$, $n^2_p = .363$. However, performance at T2 did not differ from performance at T3 ($M = 2.14$, $SD = 1.09$), $F(1, 93) = 0.95$, $p = .322$, $n^2_p = .010$. Nor did performance differ between T3 and T4 ($M = 2.03$, $SD = 1.04$), $F(1, 93) = 1.44$, $p = .233$, $n^2_p = .015$.

The main effect of constraint was not significant, $F(1, 95) = .513$, $p = .475$, $n^2_p = .005$. The main effect of WMC on performance was also not significant $F(1, 95) = .733$, $p = .394$, $n^2_p = .008$. However, there was a significant positive main effect of numeracy on performance, $F(1, 95) = 28.80$, $p < .001$, $n^2_p = .233$. No two-way interactions were significant. Neither DSE and constraint, $F(2.59, 245.80) = 0.69$, $p = .538$, $n^2_p = .007$; DSE and WMC, $F(2.59, 245.80) = 1.89$, $p = .140$, $n^2_p = .020$; nor constraint and WMC, $F(1, 95) = .467$, $p = .685$, $n^2_p = .005$ had significant two-way effects on performance. The three way interaction between DSE, constraint, and WMC on performance was not significant, $F(2.59, 245.80) = 0.89$, $p = .435$, $n^2_p = .009$.

Conflict and Working Memory Engagement. A binary logistic generalised linear mixed model was used to examine the relationship between conflict and working memory engagement. Repeated observations within participant for item and conflict factor formed a naturally nested data structure which would violate the assumption of independence

required for traditional analyses (e.g., ANOVA), therefore, multilevel modelling was used (Peugh, 2010). Item was nested within test point, and test point within participant. The dependent variable was a binary score (correct/incorrect), and the three predictors were constraint (low or high load), DSE (T1, T2, T3, and T4), and conflict factor. Working memory capacity was included as a covariate because the impact of the constraint was expected to be greater for those with lower working memory capacity than for those with higher working memory capacity. Numeracy was also included as a covariate because people with higher numeracy were expected to be further along the trajectory of learning, which may affect the point at which we would expect working memory to be engaged.

Descriptive analyses revealed that participants gave correct responses on 65.91% of the items. Participants in the low load constraint condition gave correct responses on 65.4% of the items, whereas participants in the high load constraint condition gave correct responses on 66.7% of the items. The model revealed several significant effects; however, constraint had no significant main effect, nor did it contribute to significant two- or three-way effects. These outcomes are reported in Table 2. Conflict factor had a significant negative main effect on performance, $OR=.958$ [$CI=.938, .978$], indicating that increased conflict was associated with reduced performance. DSE had a significant positive effect, indicating that performance increased with training. Follow-up pairwise comparisons revealed participants were more likely to give a correct response at T2 than T1 $OR=4.084$ [$CI=2.203, 7.571$; $p<.001$]; but no more likely to give a correct response at T3 than T2 $OR=.706$ [$CI=.367, 1.360$; $p=.297$]; and no more likely to give a correct response at T4 than T3 $OR=1.628$ [$CI=.906, 2.926$; $p=.103$]. Numeracy had a significant positive effect on performance, $OR=1.905$ [$CI=1.673, 2.168$; $p<.001$], indicating higher numeracy was associated with higher performance. Working memory capacity had a significant positive effect on performance, $OR=1.039$ [$CI=1.008, 1.071$; $p=.014$], indicating that higher working memory capacity was associated with greater performance.

Table 2. *Generalised Linear Model with Conflict Factor, DSE, Constraint, Numeracy and Working Memory Capacity on Performance.*

Source	Df _{Source}	<i>F</i>	n_p^2	<i>p</i>
Corrected Model	17	11.11	0.141	<.001
Constraint	1	0.59	0.001	.445
Conflict	1	64.11	0.053	<.001
DSE	3	14.14	0.036	<.001
Constraint*Conflict	1	1.26	0.001	.262
Constraint*DSE	3	0.50	0.001	.680
Conflict*DSE	3	5.44	0.014	.001
Conflict*DSE*Constraint	3	2.11	0.005	.098
Numeracy	1	95.11	0.076	<.001
Working Memory Capacity	1	6.02	0.005	.014

Note. $Df_{Error}=1150$

There was a significant interaction between conflict factor and DSE on performance (see Table 2). Increasing conflict predicted poorer probability of a correct response and this effect was stronger at T2, T3 and T4 than at T1 (see Figure 3). This indicates that the detrimental effect of conflict on performance strengthened as a function of test point. As reported in Table 2, no other effects were significant.

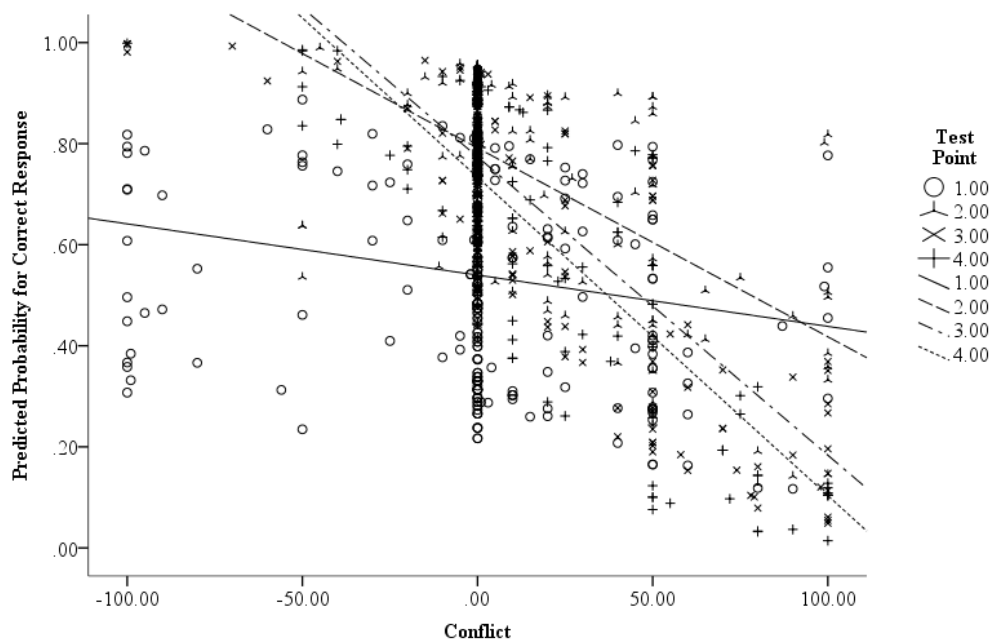


Figure 3. *Predicted probabilities for correct response as a function of conflict scores, by test point.*

Discussion.

In Study 1, training improved performance between T1 and T2. After this point, performance remained constant, regardless of cognitive load. There was a significant positive interaction between conflict and DSE on performance, such that as training increased, conflict became a stronger predictor of poorer performance. This suggests that as the participants were being trained, their conflict detection ability was improving. Training appeared to calibrate their conflict sensitivity; they developed a more finely tuned sense of when they might be right or wrong. The training may not have been extensive enough to allow participants to automate their performance to the point of no longer requiring working memory (thus mitigating the negative impact of cognitive load on performance), as would be expected at later stages of learning. This early stage of learning appeared to be preceded by conflict calibration and cognitive unease.

In contrast to our prediction and Purcell et al.'s (n.d.⁷) previous finding, there was

⁷ Paper 1 in the current thesis.

no interactive effect of DSE and load on performance. This could be due to the training not being extensive enough for automation to occur. Alternatively, it could be that the matrix load task was not hard enough to prevent participants using their working memory to assist with the main task, or it was not undertaken properly. Due to the online nature of the study, it is not possible to determine whether the participants were using memory aids. Study 2 avoids this potential issue by employing timing deadlines as cognitive constraints, which are harder for participants to circumvent with external memory aids.

Study 2

Study 2 explores a similar phenomenon to that in Study 1; it examines the relationships between cognitive constraint, DSE, conflict, and performance on the CRT. In Study 2, however, a two-response paradigm is employed such that timing restrictions were used as cognitive constraints. This paradigm allowed for a fully within-subject manipulation of training and constraint. Numeracy was measured to control for differences in the rate of automaticity (i.e. reduced effect of constraint due to practice). The within-subject manipulation of constraint reduced the need to control for individual differences, however, for consistency with and comparison to Study 1, the same individual difference measures were obtained.

Similar to the hypotheses for Study 1, we expected that DSE (test point) and cognitive constraint (response) would interact to affect performance on the CRT. That is, we expected more answer changes to occur at the intermediate points of training (T2 and T3) than at the initial test point (T1; before training) and at the final test point (T4; after automation; Purcell et al., n.d.⁸). The logical intuition model proposes that conflict may be associated with working memory engagement (De Neys, 2012), additionally, answer changes from first (timed) to second (untimed) response has been interpreted as an

⁸ Paper 1 in the current thesis.

indication of working memory engagement (Thompson et al., 2011). Therefore, we expected that higher self-reported conflict would predict a greater chance of working memory engagement as indicated by the participants' answer change from Response 1 to Response 2.

Method

Participants and Design. We used a 4 (test point) x 2 (response) within-subject design. Participants were 125 Mechanical Turk workers from the United States (81.6%), India (13.6%), and other (4.8%). Participants were paid \$US6 for participating. Mechanical Turk's exclusion capabilities were used to ensure that we were sampling high quality participants with a 95% or better performance record on previous Mechanical Turk studies. The experiment was run via the Qualtrics online survey platform. The Qualtrics exclusion capabilities were used to prevent participants participating more than once based on IP address. Study 2 was conducted after Study 1, and as participants' MTurk IDs were cross-referenced with the IDs from Study 1, those who had completed Study 1 were advised that they were ineligible to participate and directed to the end of the study. Participants were 51.2% female and 48.8% male with ages ranging from 20 to 69 ($M=38.87$, $SD=10.04$). All participants had a minimum of high school level mathematics to be eligible for the study, of these participants, 49.5% had completed a Bachelor's/Associate/Postgraduate Degree and 20.7% had partially completed a Bachelor Degree or similar.

Materials. As in Study 1, Study 2 measured Mathematical Anxiety, Numeracy, Intelligence, and Working Memory Capacity using the AMAS, BNT, RPM, and OSPAN respectively (see Study 1, Materials).

Maths Problems. As in Study 1, Study 2 contained seven blocks of conflict and no-conflict items. All items and feedback were the same as those used in Study 1. However, the presentation and response format differed. Each item was presented twice; first, under a time-limit of 5000ms (Response 1) and second, with no time-limit (Response 2). Unlike

Study 1 which employed a free response format, Study 2 used a multiple-choice format. This was to prevent typing speed from affecting whether people would respond within the time-limit imposed. Participants selected from four choices; the order of choices was randomised. See Figure 4 for an example.

Cognitive Constraint. A time-limit was imposed on Response 1 as a cognitive constraint. This type of timing constraint has been used in previous studies to reduce the likelihood of effortful thinking (e.g., Bago & De Neys, 2019). Importantly, the questions were designed such that the word length of each item was consistent with the original CRT. Items reflecting the original CRT item 1 had 23 words (+/-1) in 3 sentences, items reflecting the original CRT item 2 had the 21 words (+/-1) in 2 sentences, and items reflecting the original CRT item 3 had 44 words (+/-1) in 4 sentences. To reduce the effects of reading speed, the sentences in each question were revealed gradually until the full question was presented. Each sentence was revealed for 3000ms before the next sentence was added. The final sentence was presented at the same time as the multiple-choice options. The full question and options remained on screen for 5000ms. See Figure 4.

Procedure

Participation was completed online via Qualtrics. The general procedure followed that of Study 1, however, in contrast to Study 1, a two-response paradigm was employed in the test blocks (e.g., Thompson et al., 2013). To prepare the participant for the time-restrictions, a fixation cross was presented first for 3000ms. Immediately after providing Response 1, participants were asked: “Did you give the first answer that came to mind?” and “How confident are you in your response? (0= not at all confident, 100= absolutely confident)”. The maths problem was then presented again (Response 2), this time the full question and options were presented at once (as in Figure 4, screen D), but with no time-limit imposed. After providing Response 2, participants were asked how confident they

were in this response. A sample procedure for a conflict item is presented in Figure 4.

Training blocks were identical to those in Study 1—they included feedback and guidance and no cognitive constraints (see Supplementary Material).

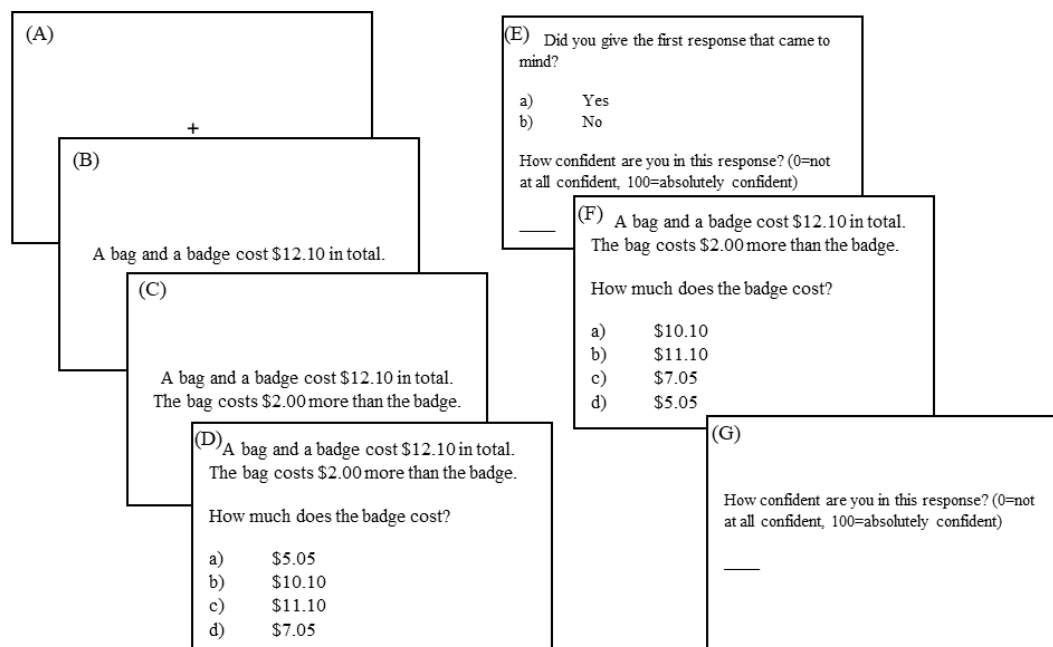


Figure 4. *Presentation of question elements for Response 1. Screens (A), (B), and (C) were shown for 3000ms each. Screen (D) was shown for 5000ms. Screens (E), (F), and (G) did not have deadlines, participants could move forward by clicking the ‘Next’ button.*

Results

Exclusion and Descriptive Analyses. No participants failed the attention check item ($N=125$). Participants who scored more than 2 standard deviations below the mean for the working memory capacity test were excluded from analysis. Consequently, 121 of the 125 participants remained in the dataset for analysis. Participants’ performance on numeracy, mathematical anxiety, intelligence, and working memory capacity is reported in Table 3, along with the bivariate correlations between each measure. The pattern of performance for Response 1 and Response 2 over testing is shown in Figure 5.

Table 3. *Means, standard deviations, and correlations between individual difference measures for participants in Study 2.*

Measure	Correlations					
	<i>M</i>	<i>SD</i>	1.	2.	3.	4.
1. Working Memory Capacity	25.26	4.30	1			
2. Mathematical Anxiety	21.92	8.12	-.013	1		
3. Intelligence	5.59	2.87	.150	-.395***	1	
4. Numeracy	2.43	1.28	.169	-.263**	.443***	1

Notes: ** $p < .01$, *** $p < .001$

DSE, Constraint and Performance. A fully within-subjects ANCOVA was used to examine the relationship between DSE (test point), constraint (response), and performance of the conflict items. A 4 (DSE: test point) x 2 (constraint: response) design was used with numeracy included as a covariate. The main effect of DSE on performance—averaged across response and numeracy— was significant, $F(3,435.17)=32.23$, $p < .001$, $\eta_p^2=.182$. Bonferroni-adjusted pairwise comparisons revealed that performance increased from T1 ($M=1.24$, $SE=.06$) to T2 ($M=1.95$, $SE=.06$), $F(1,462.52)=71.78$, $p < .001$, $\eta_p^2=.134$, but remained constant from T2 to T3 ($M=1.91$, $SE=.06$), $F(1,472.95)=.25$, $p=.617$, $\eta_p^2=.001$, and T3 to T4 ($M=1.97$, $SE=.06$), $F(1,468.06)=.47$, $p=.494$, $\eta_p^2=.001$. The main effect of constraint on performance was also significant, $F(1, 930.68)=87.25$, $p < .001$, $\eta_p^2=.086$. Performance at Response 1 ($M=1.49$, $SE=.04$) was lower than performance at Response 2 ($M=2.04$, $SE=.04$). Numeracy had a significant positive main effect on performance, $F(1, 929.93)=138.12$, $p < .001$, $\eta_p^2=.129$. The interaction effect of constraint and DSE on performance was not significant, $F(3, 435.17)=.35$, $p=.789$, $\eta_p^2=.002$. The mean performance by DSE and constraint is depicted in Figure 4.

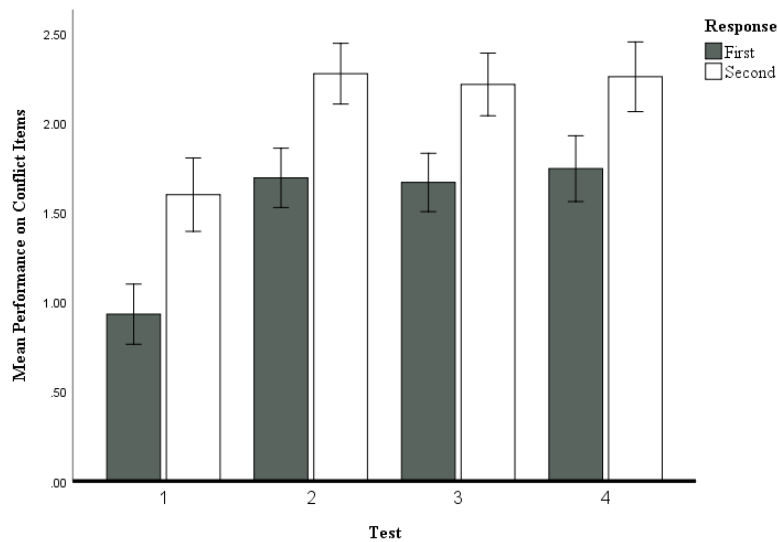


Figure 5. Mean performance on conflict items for Response 1 and Response 2 at each of the four test points. Error bars reflect $\pm 1SD$.

Conflict and Working Memory Engagement. Descriptive analyses were conducted to examine the overall trends in response changes. Items on which participants had incorrect answers on both first and second responses were coded as 00. Items on which participants had correct answers on both first and second responses were coded as 11. Items on which participants changed their answers from correct to incorrect were coded as 10. Finally, items on which participants changed their response from incorrect to correct were coded as 01. Items on which participants did not give an answer at either Response 1 or Response 2 were removed from the analysis. The frequencies of trials in each change group are presented in Table 4.

Table 4. Number of items in each change category by test point.

Test	Change Category				Total
	00	01	10	11	
1	121	51	10	99	281
2	55	45	13	187	300
3	59	47	11	185	302
4	67	58	5	58	294
Total	302	307	39	529	1177

A binary working memory engagement variable (WME) was created by collapsing

across change categories such that items on which people changed their answer from Response 1 to Response 2 (01,10) were pooled, and items on which participants gave the same answers at Response 1 and Response 2 (00,11) were pooled. Of all the trials, 29.40% fell into a ‘change’ category, indicating working memory engagement. To examine the relationship between conflict and working memory engagement, WME was used as the dependent variable in a binary logistic generalised linear mixed model. Item was nested within participant and conflict factor was entered as a predictor. Conflict factor was calculated for each item by subtracting the confidence rating (1:100) at Response 1 on the conflict item from the confidence rating (1:100) at Response 1 on the corresponding non-conflict control item. Conflict was examined as a time-varying covariate; each conflict factor was unique for each conflict item and for each participant. The model revealed a significant effect of conflict on WME, $F(1, 1048)=149.58$, $p<.001$, $OR=1.036$ [$CI=1.030, 1.042$], indicating that a greater conflict factor was associated with a higher likelihood that the participant would change their answer from Response 1 to Response 2.

Stability Index. To assess whether participants were exhibiting the same pattern of changing their answers (i.e., more frequently changing their answer from incorrect to correct [01] or always giving the correct answer [11]), we generated a stability index (SI_{change}). For each participant, the frequency of trials in each category (00,01,11,10) was calculated, the highest frequency was then recorded as a percentage of total trials and used as an index of SI_{change} . Recall that participants attempted 12 conflict items in total. Therefore, a participant who had three items ‘00’, three items ‘01’, and six items ‘11’ would have a SI_{change} of 50%. In contrast, a participant for whom all 12 items were ‘01’ would have a SI_{change} of 100%, indicating that they always responded with an incorrect answer in the cognitive constraint condition, then a correct answer without the cognitive constraint, and a participant with three items in each category (‘00’, ‘01’, ‘10’, ‘11’) would have a SI of 25%. A similar technique was employed by Bago and De Neys (2017, 2019).

The mean SI_{change} in Study 2 was 62.87% ($SD=17.87$; see Figure 6A). This mean indicates that most participants gave answers that qualified into more than one change category. The standard deviation indicates that there was a large spread of stability between participants. We also calculated a stability index for working memory engagement (SI_{WME}) by recording the responses on which each participant changed their answer (01,10)—indicating that working memory was engaged—as a percentage of their total responses. The mean SI_{WME} for Study 2 was 28.86% ($SD=20.11$; see Figure 6B). Unlike previous studies, this suggests that most people had some items on which they changed their response, indicating that working memory was engaged, and some items on which their response stayed the same – indicating that working memory was not engaged. The variability of SI_{change} and SI_{WME} are reflected in Figure 6.

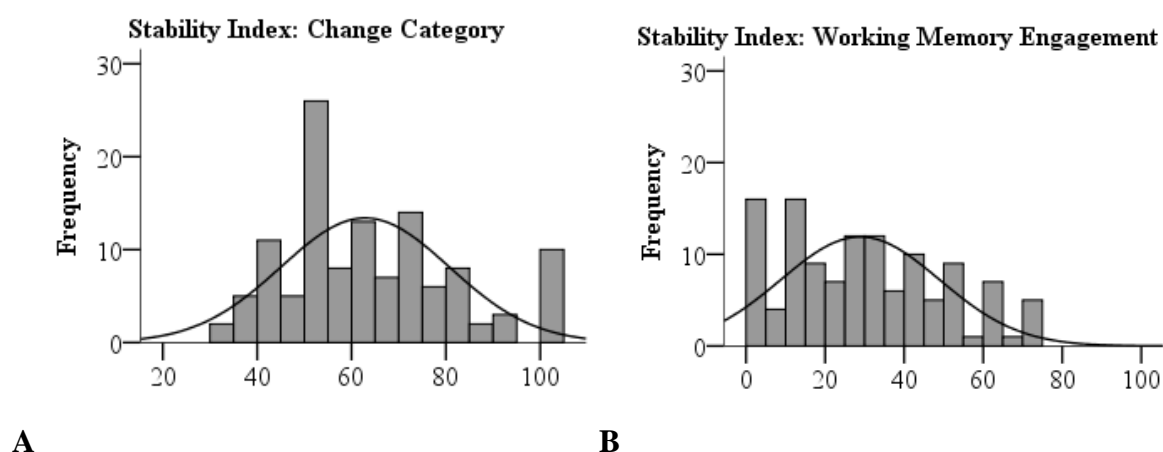


Figure 6. *Frequencies of Stability Indices for (A) Change Category (SI_{CHANGE}), and (B) WME Category (SI_{WME}).*

Discussion

In Study 2, participants' performance increased from T1 to T2 and then remained constant. Accuracy was consistently higher at Response 2 than Response 1. In contrast to our prediction, there was no interaction effect of cognitive constraint (response) and DSE (test point) on performance on the CRT. There was, however, a significant effect of conflict on working memory engagement, as measured by answer change from Response 1

[high cognitive constraint] to Response 2 [low cognitive constraint]. Higher conflict was related to a higher chance that the participant would change their answer from Response 1 to Response 2. In line with previous interpretations, this relationship between conflict and likelihood of answer change may indicate that the greater conflict a participant experienced, the more likely they were to engage working memory.

Low stability adds methodological credence to the examination of factors, like conflict, that might predict working memory engagement. Observing people both engaging and not engaging working memory lends itself to the exploration of cognitive factors that are thought to predict working memory engagement. To illustrate, consider whether the examination of a sample of highly stable participants is as meaningful as the examination of less stable participants. A high stability sample includes participants who give mostly 11s/00s or mostly 01s/10s, but few people who give both change [indicating working memory was engaged] and no-change [indicating working memory was *not* engaged] answers. Should conflict predict working memory engagement, we would expect those giving 11s/00s to have lower conflict scores than those giving 01s/10s. Comparing conflict scores and working memory engagement in a highly stable sample would naturally generate a comparison of conflict scores for people with 11s/00s versus conflict scores for people with 01s/10s. This comparison may subsequently introduce spurious factors. For example, people giving the 11s/00s may be more confident, in general, on conflict items (leading to lower conflict factors), whereas people giving the 01s/10s may be less confident on conflict items (leading to higher conflict factors). Studies with high stability samples are therefore limited in their ability to address the question of whether people switch between engaging or not engaging working memory when their perceived conflict increases or decreases.

Now consider a low stability sample in which most participants give responses that fall into more than one change category (00,01,11,10). Comparing conflict scores and

working memory engagement in a less stable sample would naturally generate a comparison of conflict scores for people with a range of patterns in changing their responses. In this instance, examining the relationship between conflict and working memory engagement allows for the consideration of whether shifts in conflict scores are associated with shifts in an individual's engagement of working memory. A study with lower stability, such as Study 2, is therefore well-placed to test the relationship between conflict and working memory engagement.

Although low stability is generally a good thing for examining intra-individual factors, the pattern of response change across training may raise some concerns. At T1 there were more "00" and less "11" trials than in T2 and T3, however, there were more "11" trials in T2 and T3 than there were in T4. The training should improve people's performance on the problems and, therefore, we would expect more "11" trials in T4. The decrease in performance is not likely to reflect systematic differences in item difficulty because the order of the items was counterbalanced. However, the decrease may be due to fatigue or boredom effects. A previous DSE study that included 18 training items and 9 test items showed no sign of fatigue effects (Purcell et al., n.d.⁹), whereas, the current study employed 18 training items and 24 test items. The possibility of fatigue effects could have led to deflated stability indices and should be considered in the design of future DSE experiments.

General Discussion

The research reported here aimed to examine the relationships between DSE, working memory engagement, and performance on the CRT; and, to investigate whether any observed changes in working memory engagement were associated with changes in conflict. Contrary to our hypothesis for Study 1, we did not find a three-way interaction

⁹ Paper 1 in the current thesis.

effect of DSE, constraint, and WMC on performance. Nor did we find a two-way interaction effect of DSE and constraint on performance. Similarly, for Study 2, we did not find an interaction effect of DSE and constraint on performance. These results suggest that, as DSE increased, the participants remained dependent on working memory for higher performance. However, Study 1 demonstrated a relationship between conflict and performance that strengthened over training. That is, as DSE increased, conflict became a stronger predictor of poorer performance. Moreover, Study 2 showed support for an association between conflict and working memory engagement; as conflict increased, the likelihood that the participant would engage working memory also increased. Together, these findings support the idea that conflict and a sense of uncertainty may be a precursor to learning.

The relationships between DSE, working memory engagement, working memory capacity demonstrated in the present studies deviated from previous findings (Purcell et al., n.d.¹⁰). Purcell et al.'s study showed a three-way interaction between these factors and performance; they demonstrated a decrease in the impact of the cognitive constraint as training increased and that this effect was greater for people with lower working memory capacity. However, although performance increased with training, the present studies did not show decreased impacts of cognitive constraints. That is, the current studies did not demonstrate that the participants had learnt the necessary procedures to the point of automation, where cognitive constraints are not expected to impact the reasoning process. While Purcell et al.'s study used a training paradigm with a matrix load task similar to that employed in Study 1, their study was conducted face-to-face with an undergraduate university sample rather than as an online procedure with a general-population sample.

These differences in environment and sample may account for the disparate

¹⁰ Paper 1 in the current thesis.

findings. It could be, for example, that the environment, face-to-face versus online, may have had an impact on the participants' motivation to learn. Conducting the study with an experimenter and fellow participants present [as in Purcell et al. (n.d.¹¹)] may have increased the participants' motivation relative to completing the study online [as in the current studies]. This may have enabled the participants in Purcell et al.'s study to learn the skill to the point of automation within four training sessions. Online participants may have been less motivated to learn quickly. Another plausible reason for different results is that the time since the general population (used in the current studies) had undertaken mathematical training (e.g., at high school or university), was longer than the time since the undergraduate sample had received mathematical training. Therefore, it is plausible that the training paradigm was not as effective for the general population because it may take them longer to 'relearn' the techniques necessary for solving the CRT.

In Study 1, greater conflict was related to reduced performance, and this relationship became stronger with training. When giving incorrect responses, the participants became increasingly aware that there may be something wrong with their response. From a logical intuition perspective, this result may indicate that, as training increased, the difference between the activation of the reasoners' heuristic processes and the activation of the reasoners' logical processes decreased. That is, even when the heuristic option continued to "win out" and the incorrect response was provided, the activation of the logical response may have been increasing as result of training. To illustrate, imagine there are two Type 1 processes that are triggered by the problem at hand (e.g., the bat and ball problem), one leading to the heuristic response (e.g., 10c) and another leading to the logical response (e.g., 5c). During the early stages of learning, the heuristic process maintains the greatest relative activation and the subject provides the heuristic

¹¹ Paper 1 in the current thesis.

response. With enough training, one would expect the logical process to obtain the greater activation relative to the heuristic process, and the subject to provide the correct response. The changes in the lead up to this switch are particularly relevant to the proposition in the logical intuition model: that increased conflict is associated with working memory engagement. This process is depicted in Figure 7.

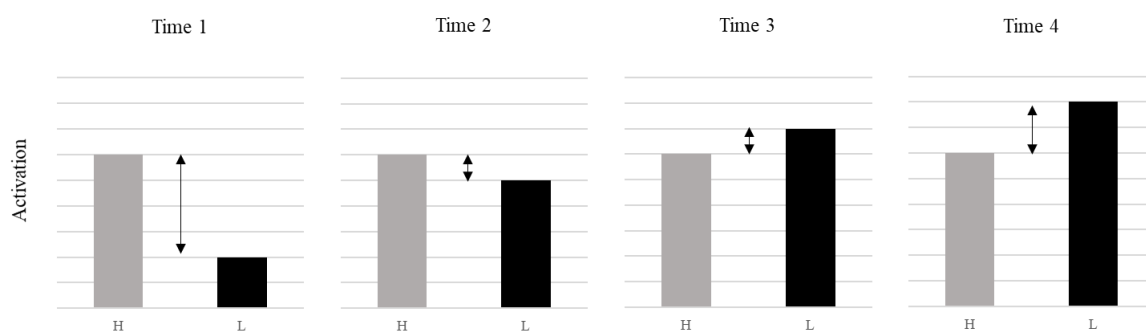


Figure 7. *Conjectural model of the increasing activation potential of a logical process across training and the interaction with thinking Type. H=Heuristic Process, L=Logical Process (adapted from Purcell et al., n.d)).*

The relationship demonstrated between conflict, DSE and performance in Study 1 supports the idea that closer relative activation of two separate Type 1 processes, induced by training, may generate conflict. Prior to the logical process gaining the greatest relative activation, the activation of the logical process may increase and become closer to that of the heuristic process. When the two processes have similar levels of activation, the logical intuition model suggests, the conflict experienced by the reasoner should be greater. Hence, in the lead up to the logical process gaining greater strength than the heuristic process, we would expect greater conflict. Study 1 found that, as training increased, incorrect responses were associated with greater levels of conflict. This finding lends support for the logical intuition model's assertion that competition between Type 1 processes, in this case induced via training, may generate conflict. Study 1, however, did not demonstrate an effect of cognitive load, and therefore, cannot speak to the relationship between conflict and working memory engagement.

Study 2, however, lends support to the logical intuition model's assertion that conflict may be associated with working memory engagement. The study demonstrated that the greater the conflict, the more likely the cognitive constraint would be effective, indicating a higher likelihood of engaging working memory. Compared to previous conflict studies, the current training paradigm effectively decreased the individual stability indices and increased the prevalence of WME. This increases the rigor of the study as a within-subject design and the strength of the interpretation that conflict and WME are associated. Despite the current evidence for the association between conflict and WME, some important queries about the nature of that relationship remain.

The logical intuition model asserts that conflict is generated by competition between implicit Type 1 processes. However, the methods used to assess conflict, both in the current article and in most previous conflict studies, employ some form of explicit Type 2 reasoning such as asking participants to consider and rate their level of confidence in their answers. Although studies have found evidence for conflict using implicit measures such as mouse-tracking (Travers et al., 2016) and skin conductance responses (De Neys, Moyens, & Ansteenwegen, 2010), these techniques have not been used in conjunction with cognitive constraint manipulations that would allow for the testing of the relationship between implicit conflict and WME. Future studies should verify whether the relationship between explicit conflict and WME, demonstrated in the current article, is mirrored by the relationship between implicit measures of conflict and WME.

The findings in the current article demonstrate that logical solutions can be reached whilst the reasoner is under cognitive constraint conditions which supports the logical intuition model's proposal that Type 1 processes can include logical principles. The findings also suggest that there is a relationship between conflict and working memory engagement which lends credence to the logical intuition model's hypothesis that conflict may be involved in the engagement of Type 2 thinking. Situations in which ambiguity is

present and learning can occur are often associated with Type 2—working memory dependent—processes. Until recently, the cognitive discomfort often experienced in situations of learning or deep contemplation has been considered as little more than an inconvenient side-effect. Here we find empirical support for the idea that cognitive unease may in fact be related to the engagement of working memory, which facilitates learning and contemplation. Cognitive conflict may be less of an inconvenient side-effect and more of a necessary precursor to reasoning.

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*, 1–43.
<https://doi.org/10.1080/13546783.2018.1507949>
- Bago, B., Raelison, 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>
- Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, 14(1), 12–23. <https://doi.org/10.1037/0096-1523.14.1.12>
- Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory and Cognition*, 38(2), 186–196. <https://doi.org/10.3758/MC.38.2.186>
- Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a Short Form. *Educational and Psychological Measurement*, 58(3), 382–398.
<https://doi.org/10.1177/0013164498058003002>
- Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment & Decision Making*, 7(1), 25–47. Retrieved from
<http://simsrad.net.ocs.mq.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=71325917&site=ehost-live>
- De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. Retrieved from

<http://pps.sagepub.com>

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., 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., & Feremans, V. (2013). Development of heuristic bias detection in elementary school. *Developmental Psychology*, 49(2), 258–269.

<https://doi.org/10.1037/a0028320>

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., Moyens, E., & Ansteenwegen, D. V. (2010). Feeling we're biased:

Autonomic arousal and reasoning conflict. *Cognitive, Affective and Behavioral Neuroscience*, 10(2), 208–216. <https://doi.org/10.3758/CABN.10.2.208>

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity:

cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, 20(2), 269–273.

<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>

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W.

(2015). Shortened complex span tasks can reliably measure working memory

capacity. *Memory & Cognition*, 43(2), 226–236. [https://doi.org/10.3758/s13421-014-](https://doi.org/10.3758/s13421-014-0461-7)

0461-7

- 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>
- Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The Abbreviated Math Anxiety Scale (AMAS). *Assessment*, 10(2), 178–182. <https://doi.org/10.1177/1073191103010002008>
- Johnson, E. D., Tubau, E., & De Neys, W. (2014). The unbearable burden of executive load on cognitive reflection: A validation of dual process theory The unbearable burden of executive load on cognitive reflection: A validation of dual process theory. *Proceedings of the Annual Meeting of the Cognitive Science Society*, (36), 36.
- Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, 164, 56–64. <https://doi.org/10.1016/j.actpsy.2015.12.008>
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
- Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(7), 1154–1170. <https://doi.org/10.1037/xlm0000372>
- Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, 124(1), 101–106. <https://doi.org/10.1016/j.cognition.2012.04.004>
- 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>

- Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review*, 19(3), 528–534.
<https://doi.org/10.3758/s13423-012-0249-3>
- Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, 48(1), 85–112. <https://doi.org/10.1016/j.jsp.2009.09.002>
- Primi, C., Donati, M. A., Chiesi, F., & Morsanyi, K. (2018). Are there gender differences in cognitive reflection? Invariance and differences related to mathematics. *Thinking & Reasoning*, 24(2), 258–279. <https://doi.org/10.1080/13546783.2017.1387606>
- Purcell, Z. A., Wastell, C. A., & Sweller, N. (n.d.). Domain-Specific Experience and Dual-Process Thinking. [Paper 1 in the current thesis]
- Raoelison, M., & De Neys, W. (2019). *Do we de-bias ourselves?: The impact of repeated presentation on the bat-and-ball problem*. *Judgment and Decision Making* (Vol. 14).
<https://doi.org/10.17605/OSF.IO/6AEC3>
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1), 3–22. <https://doi.org/10.1037/0033-2909.119.1.3>
- Stuppelle, E. J. N., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking & Reasoning*, 19(1), 54–77. <https://doi.org/10.1080/13546783.2012.735622>
- Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, 20(2), 215–244.
<https://doi.org/10.1080/13546783.2013.869763>
- 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>

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

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>

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, 28(2), 127–154. [https://doi.org/10.1016/0749-596X\(89\)90040-5](https://doi.org/10.1016/0749-596X(89)90040-5)

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37(3), 498–505.

<https://doi.org/10.3758/BF03192720>

Supplementary Material

Table 1. *Examples of training items and feedback.*

Training Item	Incorrect Feedback
A pen and a notebook cost \$25 in total. The notebook costs \$5 more than the pen. How much does the pen cost?	<p>* Sorry, your answer was incorrect. Don't worry, we'll help you find the correct solution!</p> <p>The question stated: <i>A pen and a notebook cost \$25 in total. The notebook costs \$5 more than the pen. How much does the pen cost?</i></p> <p>It's often helpful to breakdown these problems into an algebraic format. Try and give it go!</p> <p>The question states "A pen and a notebook cost \$25 in total."</p> <p>How can we write this algebraically?</p> <p>Hint: let p represent "pen" and n represent "notebook"</p> <p>a) $p + n = 25$ b) $p - n = 25$ c) $p + 25 = n$</p>
If it takes 6 adults 6 minutes to blow up 6 balloons, how long would it take 12 adults to blow up 12 balloons?	<p>Sorry! Your response was incorrect.</p> <p>The original question stated: <i>If it takes 6 adults 6 minutes to blow up 6 balloons, how long would it take 12 adults to blow up 12 balloons?</i></p> <p>The correct answer is 6 minutes. Each person can blow up ONE balloon in 6 minutes, you have 12 adults now and over 6 minutes each of them blows up ONE balloon. There are 12 adults, so they can blow up 12 balloons in total.</p> <p>Try the next question.</p>
In the desert, there is an anthill. Every day, the anthill doubles in height. If it takes 4 days for the anthill to reach 2 meters tall, long would it take for the anthill to reach 1 meter tall?	<p>Sorry! Your response was incorrect.</p> <p>The question stated: <i>In the desert, there is an anthill. Every day, the anthill doubles in height. If it takes 4 days for the anthill to reach 2 meters tall, long would it take for the anthill to reach 1 meter tall?</i></p> <p>The correct answer is 3 days. If on the third day the anthill is 1 meter tall, and it doubles in height every day, the anthill will be 2 meters tall on the fourth day.</p> <p>See if you can get the next one!</p>

Notes. *Feedback for training items related to CRT item 1 include several algebraic steps. For brevity, they are not included but are available from the authors on request.

Table 2. Means, standard deviations and correlations between measures for participants in Study 1.

Measure	<i>M</i>	<i>SD</i>	Correlations							
			1.	2.	3.	4.	5.	6.	7.	8.
1. Working Memory Capacity	21.76	5.06	1							
2. Mathematical Anxiety	22.24	8.32	-.149	1						
3. Intelligence	5.14	2.94	.280**	-.235*	1					
4. Numeracy	2.58	1.24	.295**	-.391***	.506***	1				
5. Performance T1	1.55	1.11	.266**	-.418***	.519***	.491***	1			
6. Performance T2	2.19	.93	.199*	-.237**	.319**	.438***	.630***	1		
7. Performance T3	2.14	1.09	.233*	-.350**	.472***	.439***	.684***	.790***	1	
8. Performance T4	2.03	1.04	.080	-.328***	.332**	.433***	.501***	.685***	.673***	1

Notes: * $p < .05$, ** $p < .01$, *** $p < .001$

Paper 3

Strategy and conflict on the Cognitive Reflections

Test: An eye tracking study.

Zoe A. Purcell¹, Stephanie Howarth², Colin A. Wastell¹, and Naomi Sweller¹

¹Department of Psychology, Macquarie University, Sydney, Australia

²Department of Cognitive Science, Macquarie University, Sydney, Australia

This paper has not been submitted for publication.

Abstract

To examine the nature of intuitive and reflective thinking many studies have employed the Cognitive Reflection Test (CRT). Corrective dual process models such as the default-intervention theory posit that correct reasoners typically produce a fast, intuitive response (i.e. Type 1) which can be overridden and corrected by a slow, reflective response (i.e. Type 2). In contrast, the logical intuition model proposes that correct responses on the CRT can be triggered and enacted at a Type 1 level of processing but may induce cognitive conflict. However, few studies have examined the processes underlying the CRT, and those that have, have yielded inconsistent results. To clarify which strategies are used on the CRT and whether cognitive conflict is occurring, we examined both explicit (confidence-based) and implicit (gaze-based) measures. We found no evidence for different strategies employed by correct and incorrect responders. We did, however, observe both explicit and implicit conflict for participants completing the CRT. These findings were more consistent with the logical intuition model than the default-intervention dual process interpretation of the processes underlying performance on the CRT. This study clarified previously inconsistent findings and, accordingly, several suggestions are put forward for future studies examining strategy and conflict on reasoning tasks.

Strategy and conflict on the Cognitive Reflections Test: An eye tracking study.

Dual process theories of reasoning distinguish between intuitive (also referred to as implicit or Type 1) and deliberative (also referred to as reflective or Type 2) thinking. There are various dual process models including, but not limited to, the default-intervention model (Evans & Stanovich, 2013; Kahneman, 2011), the parallel-competitive model (Epstein, 1994; Sloman, 2014; Sloman, 1996), the parallel processing model (Handley & Trippas, 2015) and hybrid models such as the logical-intuition model (De Neys, 2012, 2014), the three-stage model (Pennycook, Fugelsang, & Koehler, 2015) and the metacognitive model (Thompson, Turner, & Pennycook, 2011). The current article examines aspects of the default-intervention and logical-intuition dual process models pertaining to the strategies employed in attempts to solve the cognitive reflection test (CRT; Frederick, 2005) and cognitive conflict that may be experienced therein. The following sections outline previous research that has explored strategies and conflict on the CRT.

Part 1: Strategies for solving the CRT

The CRT has been the focus of many empirical investigations of human reasoning. The most famous item, known as the bat-and-ball problem, 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 answer that comes to mind for most respondents is 10 cents (Bago, Raelison, & De Neys, 2019; Frederick, 2005). As the name suggests, the CRT was originally developed as a measure of reflection, that is, one's ability to suppress a tempting but incorrect answer (we will refer to this as the heuristic response; e.g., 10 cents) and substitute it with the correct response (e.g., 5 cents). Accordingly, the CRT has become an exemplar of the default-intervention dual process theory of reasoning, which suggests that respondents typically produce an intuitive

response and those able to reach the correct solution do so reflectively, that is, by engaging higher level processing to override the heuristic and produce the correct response. An alternative account of CRT strategy applies the logical intuition dual process position (De Neys, 2012, 2014). Therein, responders are thought to activate two intuitive processes, one leading to the correct response and the other to the heuristic response. The process with the strongest activation “wins out”, yet, the reasoner is sensitive to the competition occurring between the two intuitive processes. However, studies examining the strategies employed on the CRT have found evidence that can support either the default-intervention or the logical intuition interpretation.

To examine the strategies employed by reasoners when solving the CRT, researchers have used paradigms that involve computer-mouse tracking (Travers, Rolison, & Feeney, 2016), think aloud protocols (Szasz, Szollosi, Palfi, & Aczel, 2017), and cognitive constraint manipulations (e.g., Bago & De Neys, 2017). Travers and colleagues (2016) examined the cursor-trajectories of participants solving the CRT. In line with the default-intervention strategy, they found that correct-respondents were tempted by the heuristic response before selecting the correct option, however, heuristic-respondents were not tempted by the correct response prior to selecting the heuristic option. Although the use of mouse-tracking has been used to examine decision strategies for other reasoning tasks (Szasz, Palfi, Szollosi, Kieslich, & Aczel, 2018), there is also evidence that such paradigms may not capture the automatic or implicit processes involved (Glöckner & Betsch, 2008; Glöckner & Herbold, 2011). Moreover, while the study by Travers and colleagues suggests that the average correct reasoner may use this strategy, the aggregation of data across participants prevented researchers establishing whether some participants may have been able to reach the correct response without considering the heuristic option first.

Szaszi, Szollosi, Palfi and Aczel (2017) examined strategies employed on the CRT with a think-aloud protocol. The participants verbally conveyed their reasoning to the experimenter as they completed the task. Reasoning strategies were categorised into ‘correct-start’ if the participant began thinking about the correct option, or ‘incorrect-start’ if the participant began thinking about the heuristic option. In the study, participants gave the correct response (10 cents) on 28% of trials. Of these trials, only 23% began with an incorrect-start, as the default-intervention account would predict. The majority (77%) of correct responses began with a correct-start thought process. Participants gave the heuristic response (5 cents) on 60% of trials, and, of these, showed evidence of deliberation on 29%. These results suggest that, while some correct-respondents displayed a default-intervention strategy, many did not. Moreover, heuristic-respondents showed evidence of engaging higher-level processes but did not override the heuristic response. Whilst these findings present a substantial challenge to the default-intervention interpretation, think aloud protocols only allow for the analysis of explicit cognitive strategies, that is, those we have conscious access to (Crutcher, 1994). In the current study we used eye-tracking to determine if the differences in strategies observed using techniques that capture explicit reasoning processes, such as mouse-tracking and think-aloud protocols, might also be observed using techniques thought to capture more implicit reasoning processes.

Cognitive constraint studies suggest that some reasoners may be able to reach the correct solution on the CRT using Type 1 processes, that is, under conditions that limit one’s ability to engage working memory resources. For example, Thompson and colleagues (2013, 2011) showed that some participants were able to successfully complete the CRT while under timing-restrictions and when instructed to provide the first response that comes to mind. Others have demonstrated that correct responding can occur

when participants are simultaneously required to memorise a dot-matrix pattern (Johnson, Tubau, & De Neys, 2016). Correct responding has also occurred when participants are under both time and load constraints (Bago & De Neys, 2019). The ability of reasoners to successfully complete the CRT under cognitive constraints has been interpreted as an indication that they were able to do so using predominantly intuitive processes. Hence, these studies pose a significant challenge the default-intervention interpretation of strategies employed on the CRT and suggest that implicit processes may play a more important role in correct-response strategies than previously thought.

Some authors have suggested that reasoners may rely on logical intuitions such that they can reach the correct solutions on the CRT without deliberation. For example, based on the Fuzzy-Trace theory (Reyna, Nelson, Han, & Dieckmann, 2009), Peters (2012) suggested that those with higher numeracy may possess superior numerical intuitions. In Szaszi and colleagues' (2017) think-aloud paradigm they had hypothesised that the two thought processes (incorrect-start and correct-start) leading to correct responses would differ as a function of numeracy. That is, higher numeracy would increase the likelihood of correct-start thought processes. However, the data were not sensitive enough to yield a definitive conclusion, possibly due to a ceiling effect on the numeracy scale. A study that employed a training manipulation of numeracy observed increased performance on the CRT and decreased dependence on working memory processes (i.e. automation; Purcell, Wastell, & Sweller, n.d.¹²). Therefore, we postulated that, as a reasoner's mathematical training increases, their strategies may shift from intuitive (when respondents typically provide the incorrect heuristic response), to reflective strategies at intermediate points (when respondents are learning the correct

¹² Paper 1 in the current thesis.

solution processes), and return to intuitive strategies at later points (when the solution process is practiced to the point of automation). In the current study we employed a training paradigm in conjunction with gaze-based measures of strategy to examine whether strategies on the CRT are moderated by numeric experience (operationalised by training) and working memory engagement.

Part 2: Conflict and the CRT

Many authors in the thinking and reasoning field have suggested that cognitive conflict may play a role in triggering deliberate thinking (e.g., De Neys, 2012, 2014; Thompson et al., 2011; Thompson & Morsanyi, 2012). That is, despite providing an erroneous response, reasoners may – at some level – register the fact that their answer is not completely warranted. Thompson and colleagues have related this idea to the metacognitive memory literature and 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. Alternatively, though not necessarily incompatible, the logical intuition model asserts that multiple Type 1 processes can be initiated at once, and that their relative activation can elicit conflict. When the initiated processes have similar levels of activation, conflict is generated, and deliberative thought is engaged. Evidence that heuristic respondents may be sensitive to the logical solution and experience disfluency or conflict comes from a growing area of research on conflict detection.

Conflict detection studies typically compare ‘lure’ and ‘no lure’ versions of bias tasks, such as base-rate or syllogistic reasoning problems (Bago & De Neys, 2017). Lure items in conflict studies examining the CRT reflect the original CRT in that there is a heuristic but incorrect response, compared to no lure items that have no heuristic incorrect

response. For example, the bat and ball problem introduced earlier has the heuristic lure of 10 cents, in contrast, a no lure version 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 cents. Heuristic responding on lure items has been compared to correct responding on no lure items to determine whether the heuristic respondents are sensitive to the conflicting mathematical principles in the lure items. However, conflict sensitivity has been conceptualised and operationalised in different ways across conflict studies.

Conflict sensitivity has been observed on various bias tasks using measures that range from self-report, requiring an explicit awareness of the conflict, to implicit physiological indicators of conflict (see De Neys, 2012). For example, relative to participants who respond correctly on no lure items, participants who respond heuristically to the lure versions have reported lower confidence (De Neys, Cromheeke, & Osman, 2011); shown higher activation in the anterior cingulate cortex (De Neys, Vartanian, & Goel, 2008), an area of the brain thought to mediate conflict in decision-making (Botvinick, Cohen, & Carter, 2004); and shown higher autonomic nervous system activation, indicating greater arousal (De Neys, Moyens, & Ansteenwegen, 2010). Regarding the CRT, some studies have observed differences in explicit self-report measures, for example, relative to problems on which respondents gave the correct answer (correct-no lure items), participants displayed lower confidence (De Neys, Rossi, & Houdé, 2013) and greater feelings of error (Gangemi, Bourgeois-Gironde, & Mancini, 2015) on problems on which respondents gave the heuristic answer (heuristic-lure items). However, studies examining the more implicit registration of conflict during the CRT are fewer and contain mixed results.

Some studies have found differences between lure and no lure CRT items on

implicit indicators of conflict (e.g., Stuppel et al., 2017; Travers et al., 2016). For example, Stuppel et al., (2017) found that response latencies were greater for correct-no lure responses than heuristic-lure responses for the first CRT item (the bat and ball problem) but not the other two items. Travers et al. (2016) employed the extended eight-item CRT (Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2016) with measures of response times and cursor-trajectories. When comparing correct-no lure and heuristic-lure responses (as is typical in conflict detection studies) they did not find evidence for differences in response times, distance the cursor travelled, or the number of cursor movements. However, concerns have been raised about the rigour of mouse-tracking methodologies which have been shown to differ from more sensitive attentional measures, such as eye tracking (Ball, Lucas, Miles & Gale, 2003). It may be that with more sensitive tools, like eye tracking, implicit conflict might be detected on the CRT heuristic-lure items relative to the CRT correct-no lure items.

Conflict may be moderated by domain-specific experience. Previous studies have demonstrated that performance on the CRT can be improved via training manipulations (Purcell, Wastell, & Sweller, n.d.¹³; Purcell, Wastell, Sweller, & Howarth, n.d.¹⁴). Purcell, Wastell, Sweller, et al. (n.d.¹⁵) demonstrated that as performance increased, conflict sensitivity became more finely tuned. That is, participants' explicit confidence ratings were more predictive of performance with increased training. Relatedly, Frey and De Neys (2017) investigated whether conflict sensitivity is a domain-general phenomenon. They examined the relationship between conflict detection on several bias tasks such as base-rate neglect problems, syllogisms, and the bat and ball problem. However, they found no clear indication that conflict sensitivity on one task predicted conflict sensitivity

¹³ Paper 1 in the current thesis.

^{14, 15} Paper 2 in the current thesis.

on other tasks. This suggests that conflict sensitivity may be linked more specifically to the processes underlying each of the tasks, rather than to a general cognitive mechanism. Together, these findings suggest that conflict detection may be closely linked to the reasoner's domain-specific experience and that a reasoner's sensitivity to conflicting processes may change over training. Therefore, in addition to examining whether conflict was observed on explicit confidence-based and implicit eye-tracking measures, we explored whether explicit and implicit conflict changed with practice.

The Current Paradigm

Previously, cognitive processing on the CRT has been examined using think aloud protocols, response latencies, confidence ratings, and mouse-tracking techniques. In the current paradigm, conflict was measured using confidence ratings and two eye tracking measures. Eye tracking is a non-invasive method for examining conflict at the time of processing, without relying on subjective, explicit, post-hoc judgments or influencing the reasoning process. Eye movements have been recorded to explore the cognitive and attentional processes underlying problem solving strategies. They can provide us with information about the cognitive processes that the researcher cannot determine from performance alone and the participant may not be consciously aware of (e.g., Bruckmaier, Binder, Krauss, & Kufner, 2019; Green, Lemaire, & Dufau, 2007; Stephen, Boncoddio, Magnuson, & Dixon, 2009).

There are several eye-tracking measures that can be examined to assess cognitive processes throughout reasoning. For this study, we focussed on 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. Fixations have been used to determine which pieces of information a participant has considered and which they have not. For example, Ball, Phillips, Wade and Quayle (2006) found that participants fixated on the premises of

incongruent syllogisms (i.e. when the logical validity was incongruent with the statement's believability) more frequently than congruent syllogisms. Dwell is thought to reflect the depth of processing (Glöckner & Herbold, 2011; Velichkovsky, 2014; Velichkovsky, Rothert, Kopf, Dornhöfer, & Joos, 2002), such that longer dwell times reflect more complex processing. For example, silent reading is related to short dwell times (~225 milliseconds; Rayner, 1998), while typing is related to medium dwell times (~400 milliseconds; Rayner, 1998), and calculating weighted sums is associated with long dwell times (>500 milliseconds; Horstmann, Ahlgrimm, & Glöckner, 2009). Therefore, we measured both fixations and dwell to examine differences in the information that may be considered in the reasoning process and the depth with which that consideration occurs. The proportion of the number of fixations and the amount of dwell that occurred on the each of the four multiple choice options was used to calculate conflict. This was used as an indicator of the amount of indecision with which a participant selected a response, the greater the proportion of fixations and dwell on the non-selected responses the greater the indication of conflict.

The current study aimed to address two research questions. First, do response strategies differ for correct- and heuristic-respondents, and do these change with practice? Second, is there evidence of explicit or implicit conflict on the CRT, and does explicit or implicit conflict change over practice? To achieve these aims, a CRT training task was employed with two within-subject manipulations: test points (T1, T2, T3, T4) and problem type (lure and no lure) and a between-subjects cognitive load condition (high load and low load). Hypotheses for the behavioural outcomes, and for the two research questions are outlined below.

Behavioural. To explore the relationship between practice, CRT performance, and working memory engagement, we examined the effects of test point, a secondary

cognitive load, and working memory capacity on performance on the lure items. We expected that each of the three factors would have main effects on performance: the higher the test point, the higher the performance; the higher the load, the lower the performance; and, the higher WMC, the higher the performance. We also hypothesised a three-way interaction between test point, load condition and WMC; we expected that test point (T1, T2, T3, T4), load condition (high load or low load), and WMC would interact to affect performance on the lure problems (Purcell, Wastell, & Sweller, n.d.¹⁶; cf. Purcell, Wastell, Sweller, et al., n.d.¹⁷). That is, we expected that the effect of load condition would be greater at intermediate points of training (T2 and T3) than prior to training (T1) or after automation (T4). However, we expected that this interaction would be stronger for participants with low WMC, who were more likely to be affected by the cognitive load, than those with high WMC. Support for this hypothesis would lend credence to the assertion that automaticity (decreased working memory dependence) can occur with sufficient domain-specific practice. Improved CRT performance with practice was essential for the examination of the following two sets of hypotheses relating to strategy and conflict.

Strategy. Eye-tracking measures were examined as indicators of participants' strategies on the CRT. To study the strategies employed by correct- and heuristic-respondents on the lure items, we examined participants' fixations and the duration of those fixations (dwell) that occurred on the four multiple-choice options (see Method). In line with Travers et al. (2016), we expected that prior to training, correct-respondents would consider the heuristic option. In contrast, heuristic-respondents were not expected to consider the correct option. That is, we predicted that at T1, correct-respondents would

¹⁶ Paper 1 in the current thesis.

¹⁷ Paper 2 in the current thesis.

show greater proportions of fixations and dwell on the heuristic response than heuristic-respondents would on the correct response. Additionally, we expected that the pattern of response strategy would change over training such that the respondents would display decreasing consideration of the heuristic response and alternative options relative to the correct response. That is, we predicted that, as test point increased, the proportional number of fixations and duration of dwell on the heuristic response would decrease.

Conflict. Eye-tracking and behavioural measures were examined as indicators of conflict on the CRT. To inspect explicit and implicit conflict on the CRT we tested the impact of trial type (heuristic-lure and correct-no lure) on three conflict measures. In line with previous conflict studies, we compared conflict on heuristic-lure trials to conflict on correct-no lure trials (e.g., De Neys et al., 2013). Confidence ratings were reverse scored to reflect explicit conflict; these scores were expected to be higher for heuristic-lure items than correct-no lure trials. That is, we expected that the heuristic reasoner would have some explicit awareness that their response may be incorrect.

To examine more implicit indicators of conflict, we examined the dispersion of fixations and dwell across the four multiple choice options. Both fixations and dwell were measured because there is no previously established measure for implicit conflict and, as explained earlier, they are thought to reflect slightly different aspects of cognitive processing. A lower proportion of fixations and dwell on the chosen 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 conflict. 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, indicate lower conflict. Therefore, the proportion of fixations and dwell were expected to be more

evenly spread across the response options for heuristic-lure items than correct-no lure items.

In addition to examining the overall patterns of fixations and dwell, we also calculated a single conflict factor for each measure. An implicit conflict factor based on fixations was calculated by taking the inverse of the proportion of fixations on the response that the participant selected. Similarly, a dwell-based conflict factor was generated by taking the inverse of the proportion of dwell on the selected response. As suggested above, greater consideration of the alternatives was thought to indicate greater conflict. The number of fixations and duration of dwell on the alternative responses relative to the selected response was captured by taking the inverse of the proportion on the selected response. Therefore, both implicit conflict factors were expected to be greater for heuristic-lure items than correct-no lure items. That is, we expected that reasoners selecting the heuristic answer on lure items would demonstrate greater implicit conflict than reasoners selecting the correct answer on no lure items. Additionally, we examined the relationship between the explicit, confidence-based conflict measure, the implicit fixation-based measure, and the implicit dwell-based measure. We expected all three measures to be positively correlated; this would lend credence to the use of confidence ratings, fixations, and dwell, as indicators of cognitive conflict.

Method

Participants & Design

A 4 (test point: T1, T2, T3, T4; within-subjects) x 2 (problem type: lure and no lure; within-subjects) x 2 (load condition: low and high load; between-subjects) mixed design was used. Participants were 38 undergraduate psychology students at Macquarie University (Sydney, Australia) awarded course credit for participation. All participants had normal vision. Participants were randomly allocated to the low load (N=20) or high

load (N=18) condition. Participants were 27 females and 11 males with ages ranging from 18 to 36 ($M=19.76$, $SD=3.54$).

Apparatus & Materials

There were two components of the study, the WMC test and the reasoning and load task. For both components, participants were tested individually; all testing was conducted in the same room and the experimenter was seated behind the participant.

Working memory capacity test. The WMC test was conducted on a laptop using the Qualtrics survey platform. To assess WMC, participants completed a short version of the Operation Span Task (OSPAN; Foster et al., 2015; Unsworth, Heitz, Schrock, & Engle, 2005). The OSPAN required participants to remember and recall several letters while assessing validity of several short mathematical problems. For example, if a participant was presented with the letter “D” to recall, followed by an equation “ $(9 / 3) - 2 = 2$?”, they would respond with True if they believe the equation is valid, or False if they thought it was not valid. They would then be presented with another letter to remember and equation to assess, and so on for the length of the letter span. A recall sheet was then provided for the respondent to serially recall the letters they had seen. The test included three practice trials with letter spans of two, for which they received feedback, and then six test trials with letter spans of three to eight. Accurate serial recall was summed; scores had a possible range of 0 to 33. The OSPAN took approximately 15 minutes.

Reasoning task. The reasoning task was presented on a 24.5-inch LCD monitor (BenQ XL2540, refresh rate 240 Hz, natural resolution 1920 x 1080). 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 1000 Hz (EyeLink 1000; SR Research Ltd., Osgoode, Ontario, Canada). A chinrest was used to stabilise head movements and maintain viewing distance (800mm). Data was extracted using Eyelink Data Viewer (SR-Research) and analysed with SPSS Version 26.0.

For coding purposes, Areas of Interest (AOIs) were assigned to the multiple-choice alternatives. These were labelled relative to the response that the participant had chosen on each trial as ‘Selected’, ‘Other-relevant’, ‘Other-1’ and ‘Other-2’. Therefore, on lure items, 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 in the same way except that there was no theoretical distinction between the incorrect ‘Other’ AOIs. The analyses treat each AOI separately in order 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. Multiple-choice response options were randomly allocated to one corner of the screen for each item (see Figure 2). Therefore, AOIs were dynamic such that they reflected the option value and not the placement. The number of fixations and their duration time (dwell) was summed for each AOI by trial.

The reasoning task employed in the study is based on previous paradigms that have been used to improve performance on the CRT (Purcell, Wastell, & Sweller, n.d.¹⁸; Purcell, Wastell, Sweller, et al., n.d.¹⁹). Participants were presented with seven blocks, each with six maths problems, in the order: T1, training-block 1, T2, training-block 2, T3, training-block 3, and T4. The order of the blocks was counterbalanced such that the items

^{18, 20} Paper 1 in the current thesis.

¹⁹ Paper 2 in the current thesis.

used in T1 for one half of the participants were the same items used in T4 for the other half of the participants. The order of items was randomised within each block to prevent order effects. Each block contained six items, including three lure and three no-lure 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 3 spiders 3 minutes to make 3 webs. How long would it take 100 spiders to make 100 webs?” (correct answer: 3 minutes). This item has an incorrect heuristic response of “100”. An example of a corresponding no lure item is: “It takes 1 factory 10 days to build 20 cars. How many days would it take 1 factory to build 40 cars?” (correct answer: 20 days). This item has no heuristic response.

During test blocks only, matrix memory tasks were imposed simultaneous to the mathematical problems. Matrix memory tasks use limited cognitive resources (Bethell-Fox & Shepard, 1988). These tasks have, therefore, been used to restrict working memory engagement on reasoning tasks (e.g., Bago & De Neys, 2017; Johnson, Tubau, & De Neys, 2014; Purcell, Wastell, & Sweller, n.d.²⁰). In the present study, all participants had to memorise a matrix pattern while completing test items, however, those in the low load condition remembered simpler matrices (one-piece patterns) than those in the high load condition who remembered more complex patterns (two- or three-piece patterns; see Figure 1). Each square that was correctly coloured (or not) was scored 1. Scores are reported as percentages out of 9.

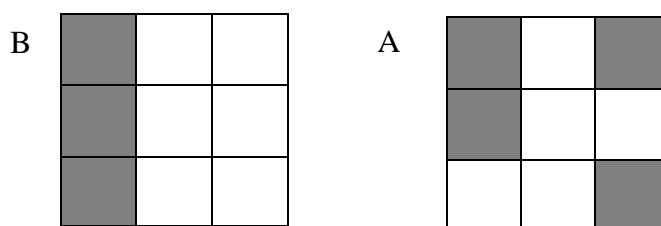


Figure 1. *Example of matrices for (A) low load and (B) high load conditions.*

For each test item, participants were first presented with a matrix pattern to remember for 900ms and the instructions: “Memorise which cells are highlighted. You will be asked to recall them soon.” The maths problem and multiple-choice alternatives were then presented (see Figure 2). The alternatives were presented in the four corners of the screen, equidistant from the centre where the question was displayed, and the position of the answers was randomised (see Figure 2). On the next screen they were presented with the question: “How confident are you in your answer?” and a slider to position from 0 (not at all confident) to 100 (absolutely confident). On the following screen, they were presented with a blank grid and the instructions: “Click the cells that were highlighted in the grid you saw earlier.” For screens that did not have auto-proceed time limits, the participant could proceed by clicking Next.

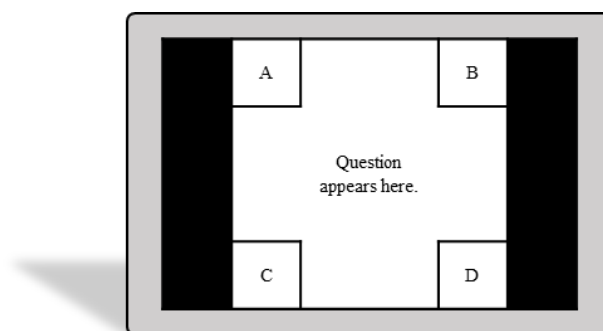


Figure 2. *Layout used for maths problems. Participants were required to choose an answer from the four options in the corners. The position of the answers was randomised for each item.*

Only training blocks included feedback (correct/incorrect) and guidance (see Appendix C). For items in the training blocks, participants were presented with a question

in the layout of Figure 2. Once they had responded, another screen appeared with feedback: “Correct” or “Incorrect”. On the following screen, those who gave correct responses were provided with a short explanation of why this was the correct response. Those who gave incorrect responses received more interactive guidance, for example, learning to translate questions into algebraic forms by following several steps. This training program has been successfully implemented to improve performance on CRT-like problems in previous studies (e.g., Purcell, Wastell, & Sweller, n.d.²¹). The reasoning task took approximately 40 minutes.

Procedure

Participants gave consent before completing the WMC test and reasoning task. The order of the WMC test and reasoning task was counterbalanced. Prior to commencing the WMC test, basic demographic details were collected (via Qualtrics). Simple computerised instructions were provided throughout the WMC test. 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. Participants were provided with a pen and notepad, but were advised that they could not use the notepad for the matrix memory tasks (notepads were checked after the task to ensure they were not used for the matrices). The chair and chinrest were then adjusted to a comfortable height for the participant and the eye-tracking procedure began.

The 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, and the reasoning task then began. Two three-minute breaks were included in which participants were advised to take

²¹ Paper 1 in the current thesis.

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. 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

The results are presented in three sections. The first section presents the behavioural results relating to performance on lure items by test point and load condition. The second section includes the analyses relating to strategies employed on the lure items and their interaction with test point; these analyses are separated for fixation-based and dwell-based measures. The third section examines conflict on the CRT by comparing heuristic-lure and correct-no lure trials. This third section is divided into four subsections examining a confidence-based measure of explicit conflict, a fixation-based measure of implicit conflict, a dwell-based measure of implicit conflict, and finally, the relationship between these measures.

Behavioural Results

We examined the interaction between test point, load condition and WMC on performance on the lure items. Performance on the matrix load tasks was high ($M=96.00$, $SD=3.75$), indicating that participants were completing the load task properly. Descriptive analyses revealed that participants gave correct responses on 75.66% of the lure items. Participants in the low load condition gave correct responses on 72.08% of the lure items, whereas participants in the high load condition gave correct responses on 79.63% of the lure items. A mixed ANOVA was used to examine the effects of test point (within-subjects: T1, T2, T3, and T4), load condition (between-subjects: low or high load) and working memory capacity (continuous) on performance (score on lure items). Where

sphericity was violated, Greenhouse-Geisser adjusted results are reported. Results of the ANOVA are reported in Table 2.

Table 2. *ANOVA with effects of Test Point, Load Condition, and Working Memory Capacity on Performance.*

Source	Df _{Source}	<i>F</i>	η^2_p	<i>p</i>
Test Point	1	45.102	0.570	<.001
Load Condition	1	.981	0.028	.329
WMC	1	7.88	0.188	.008
Test Point *Load Condition	2.138	.619	0.018	.552
Test Point *WMC	2.138	.549	0.016	.591
Load Condition*WMC	1	2.096	0.058	.157
Test Point *Load Condition* WMC	2.138	2.63	0.072	.075

Note. $Df_{Error}=31.154$

As reported in Table 2, load condition did not contribute to any main, two-way interaction or three-way interaction effects. Therefore, there is no evidence for an effect of load condition at any test point or level of working memory capacity. However, there was a significant positive main effect of WMC, indicating that as WMC increased, so too did performance. Test point also had a significant main effect on performance; pairwise comparisons between test points were examined against a Bonferroni adjusted alpha of .017. Performance significantly increased from T1 ($M=1.39$, $SD=1.00$) to T2 ($M=2.45$, $SD=.76$), $F(1,32)=57.05$, $p<.001$, $\eta^2_p=.641$, and from T2 to T3 ($M=2.74$, $SD=.64$), $F(1,32)=10.201$, $p=.003$, $\eta^2_p=.242$. However, there was a significant decrease in performance from T3 to T4 ($M=2.50$, $SD=.60$), $F(1,32)=8.94$, $p=.005$, $\eta^2_p=.218$. For additional descriptive results see Appendix B, Table A.

Strategies Employed on the CRT

Strategy Measured by Fixations

To examine the strategies employed on lure items for correct- versus heuristic-respondents, a linear mixed model was run with AOI nested within accuracy, accuracy within item, item within test point, and test point within participant. Trials on which participants gave neither the correct nor heuristic response were removed for this analysis. The AOIs were coded to reflect the ‘Selected’ response (correct for correct-respondents and heuristic for heuristic-respondents), the ‘Other-Relevant’ response (heuristic for correct-respondents and correct for heuristic-respondents), the ‘Other-1’ and ‘Other-2’ responses (corresponding to the remaining two multiple choice options). The model included four predictors: item²² (1,2,3), test point (T1, T2, T3, T4), accuracy (correct, heuristic), and AOI (Selected, Other-Relevant, Other-1, Other-2). The dependent variable was the proportion of fixations. The effects included in the model and the output of each effect is reported in the Appendix C, Table A. The three-way interaction between test point, accuracy and AOI was not significant, nor were any two-way interactions (see Appendix C, Table A). However, the model revealed a significant main effect of AOI, $F(3,1240.27)=150.53, p<.001, \eta_p^2=.267$.

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 comparisons revealed that the proportion of fixations was greatest for the ‘Selected’ AOI than all other AOIs, and that the proportion of fixations was not significantly different between any of the three ‘Other’ AOIs (see Appendix C, Table C). There was no evidence that this pattern of fixations differed between correct- and heuristic-respondents.

²² Item is used here to refer to the structure of the problem. For example, lure items structured to match the first item in the CRT (the bat and ball problem) were coded “1” along with their no lure counterparts, and so on for the second and third items. While the differences between the items were not the focus the current study, we felt it was important to control for any such differences.

Specifically, there was no evidence that correct-respondents showed greater consideration of the heuristic response than heuristic responders did of the correct response. Moreover, there was no indication that these strategies changed across test points.

Strategy Measured by Dwell

A similar analysis was conducted to examine the strategies employed on lure items for correct- and heuristic-respondents as measured by dwell. A linear mixed model was run with the same data structure and predictors as above, however, the dependent variable was the proportion of dwell occurring in each AOI (for the specific effects and outcomes see Appendix C, Table B). The model demonstrated a similar pattern of results to the analysis above, that is, the main effect of AOI was significant, $F(3,1221)=409.65$, $p<.001$, $\eta_p^2=.502$ and no two- or three-way interactions reached statistical significance. Bonferroni adjusted pairwise comparisons revealed that the proportion of dwell was greater for the ‘Selected’ response than all ‘Other’ responses and the proportion of dwell did not differ between the ‘Other’ responses (see Appendix C, Table D). For strategy measured via fixations, there was no indication that the reasoners’ strategies differed between correct- and heuristic-respondents, nor was there evidence that it differed across test points.

Conflict

To explore the patterns of explicit and implicit conflict during the reasoning task, we examined conflict for lure trials on which heuristic responses were made (heuristic-lure) and conflict for no lure trials on which correct responses were made (correct-no lure). We first examined the pattern of explicit conflict via participants’ self-reported confidence in their response. We then examined the pattern of implicit conflict, first via measures of fixations, and then via measures of dwell. Lastly, we examined the relationship between the three conflict measures, namely the explicit conflict factor, and

two newly generated implicit conflict factors.

Explicit Conflict Measured via Confidence

A linear mixed model was used to examine explicit conflict measured via reverse-coded confidence ratings. The model included the predictors: item (1,2,3), test point (T1, T2, T3, T4) and trial type (heuristic-lure, correct-no lure). The effects included in the model and their outcomes are reported in the Appendix C, Table E. The two-way interaction between test point and trial type was significant, $F(3,543)=3.20$, $p=.023$, $\eta_p^2=.017$. Therefore, simple effects were examined for trial type (heuristic-lure and correct-no lure) at each level of test point. The results were compared to a Bonferroni adjusted alpha of .0125. The differences between explicit conflict on heuristic-lure and correct-no lure trials was significant at T1, $F(1,543)=80.41$, $p<.001$, $\eta_p^2=.129$; T2, $F(1,543)=35.03$, $p<.001$, $\eta_p^2=.061$; and T4, $F(1,543)=54.56$, $p<.001$, $\eta_p^2=.091$; but not T3, $F(1,543)=5.87$, $p=.016$, $\eta_p^2=.011$ (see Figure 3).

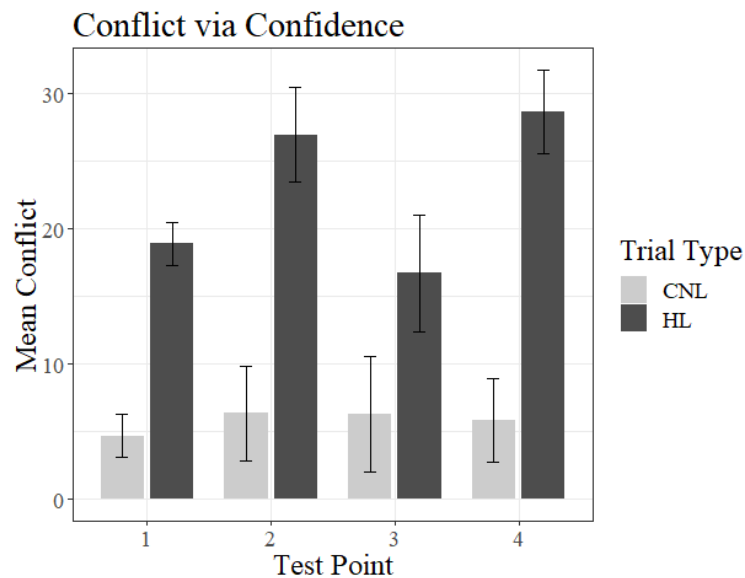


Figure 3. *Explicit conflict by test point and trial type (CNL=Correct-no lure trials; HL=Heuristic lure trials).*

Implicit Conflict Measured via Fixations

A linear mixed model was used to examine implicit conflict as measured by the

proportion of fixations that occurred on each multiple-choice option. As above, AOIs were coded to reflect the ‘Selected’, ‘Other-Relevant’, ‘Other-1’ and ‘Other-2’ response options. The predictors were item (1,2,3), test point (T1, T2, T3, T4), trial type (heuristic-lure, correct-no lure), and AOI. The dependent variable was the proportion of fixations. The effects included in the model and their outcomes are reported in the Appendix C, Table F. Notably, the three-way interaction between test point, trial type, and AOI was not significant, $F(9,2170)=1.16$, $p=.320$, $\eta_p^2=.005$. However, the two-way interaction between trial type and AOI was significant, $F(9,2170)=5.427$, $p=.001$, $\eta_p^2=.007$.

The 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 results were compared to a Bonferroni adjusted alpha of .0125. These results are presented in Figure 4A. There was a significant difference between the proportion of fixations in the ‘Selected’ AOI for heuristic-lure and correct-no lure trial types, $F(1,2170)=10.94$, $p=.001$, $\eta_p^2=.005$. The differences between the proportion of fixations on heuristic-lure and correct-no lure trial types were not significantly different for AOIs: ‘Other-relevant’, $F(1,2170)=0.38$, $p=.536$, $\eta_p^2<.001$; ‘Other-1’, $F(1,2170)=4.16$, $p=.046$, $\eta_p^2=.002$; or, ‘Other-2’, $F(1,2170)=.01$, $p=.893$, $\eta_p^2<.001$. The proportion of fixations on the selected responses were lower on the heuristic-lure than the correct-no lure trial types (see Figure 4A). This suggests that when responding with the heuristic response on lure problems the participants had considered the non-selected options to a greater extent than they had when responding correctly on the no lure problems. Therefore, it appears that implicit conflict was occurring on the heuristic-lure trials, and that this conflict might be captured by the proportion of fixations occurring on the selected response. However, there was no evidence that implicit conflict (captured by the two-way interaction between trial type and AOI) differed across test point.

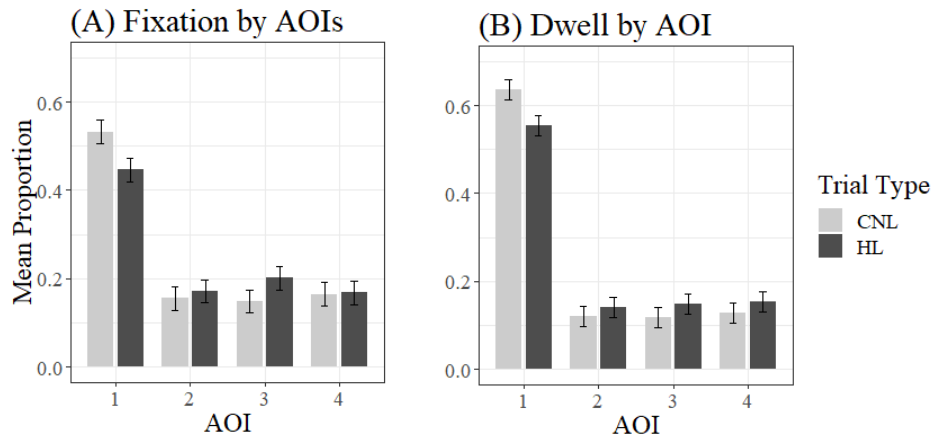


Figure 4. *Implicit conflict explored via: (A) the proportion of fixations by AOI and trial type, and (B) the proportion of dwell by AOI and trial type. CNL=Correct-no lure trials; HL=Heuristic lure trials. AOIs are coded: 1 = Selected response (correct for CNL and incorrect lure for HL), 2 = Other relevant for HL trials (i.e. the correct response) and other for CNL, 3 = Other, 4 = Other. CNL=Correct-no lure trials; HL=Heuristic lure trials.*

Implicit Conflict Measured via Dwell

A linear mixed model was used to examine implicit conflict via the proportions of dwell on the multiple-choice options. The model's predictors were: item (1,2,3), test point (T1, T2, T3, T4), trial type (heuristic-lure, correct-no lure), and AOI (Selected, Other-Relevant, Other-1, Other-2). The dependent variable was the proportion of dwell. The effects and outcomes of the model are presented in Appendix C, Table G. For conflict via fixations, the three-way interaction between test point, trial type and AOI was not significant, $F(9,2170)=.78, p=.637, \eta_p^2=.003$. However, the two-way interaction between trial type and AOI was significant, $F(9,2170)=6.07, p<.001, \eta_p^2=.008$.

The two-way interaction between problem type and AOI was examined more closely by comparing the proportion of dwell for heuristic-lure and correct-no lure trial types for each AOI. These results are seen in Figure 4B. There was a significant difference between the proportion of dwell on the AOI for the selected response for

heuristic-lure and correct-no lure trial types, $F(1,2170)=13.02$, $p<.001$, $\eta_p^2=.005$. The differences between the proportion of dwell on heuristic-lure and correct-no lure trial types were not significantly different for AOIs: ‘Other-relevant’, $F(1,2170)=.68$, $p=.410$, $\eta_p^2<.001$; ‘Other-1’, $F(1,2170)=1.81$, $p=.186$, $\eta_p^2<.001$; or, ‘Other-2’, $F(1,2170)=1.38$, $p=.238$, $\eta_p^2<.001$. This suggests that the heuristic-respondents experienced greater implicit conflict on the lure items relative to the correct-respondents on no lure items, and that this difference is captured by the proportion of dwell on the selected responses. However, there was no indication that implicit conflict via dwell differed across test point.

A Comparison of Explicit and Implicit Conflict Measures

To examine the relationships between each of the conflict measures we used the explicit conflict factor used in earlier analyses, and two new implicit conflict factors. As examined above, a confidence-based conflict factor (CF-C) was calculated by reverse-coding participants’ confidence ratings. Additionally, two new implicit conflict factors were generated as the inverse of the proportion of fixations (CF-F) and dwell (CF-D) on the participants’ selected response for each trial. We ran three bivariate correlations between these factors; CF-C was weakly correlated with CF-F ($r=.133$, $p<.001$) and CF-D ($r=.185$, $p<.001$). CF-F and CF-D were strongly correlated ($r=.874$, $p<.001$). This indicates moderate convergent validity for the three measures but also suggests that the explicit (CF-C) and implicit (CF-F and CF-D) conflict factors may reflect separate psychological phenomena.

Discussion

The main findings of the study did not indicate that strategies employed on the CRT were different for correct and heuristics respondents, nor did it demonstrate support for the suggestion that strategy changes with training. Regarding conflict, the study’s

findings support the suggestion that explicit and implicit conflict are evident for heuristic responses on lure items relative to correct responses on no lure items. However, there was no evidence to suggest that the effect of conflict changed across training. Broadly, these findings support the logical intuition interpretation of CRT performance, however, when considered alongside previous studies it is evident that a more complex model may need to be developed to account for the strategies and conflict demonstrated for reasoners completing the CRT.

Eye tracking measures were employed to compare the strategies used by correct and heuristic respondents. As expected, the analysis of both fixations and dwell found that respondents looked at the response that they had eventually selected more often and for longer than they looked at the alternatives. Unexpectedly, however, there was no evidence to suggest that the respondents who eventually selected the correct response had considered the heuristic option to a greater extent than the heuristic respondents had considered the correct response. This suggests that both correct and heuristic respondents selected their final answers with little consideration of the alternatives. Little consideration of the alternatives may indicate low deliberation. As such, this finding is in line with cognitive constraint studies that have found correct and heuristic responding under conditions that discourage deliberation (e.g., timing restrictions or cognitive loads; Bago & De Neys, 2019) and the logical intuition model's assertion that logical principles can be enacted in Type 1 processing (De Neys, 2012). However, it challenges the default-intervention interpretation of CRT strategies.

The default-intervention interpretation of CRT strategies stipulates that an initial incorrect response is generated via Type 1 processes but that this response can be overridden by Type 2 process that can then generate the correct response (Kahneman & Frederick, 2005). In contrast to the current findings, Travers et al.'s (2016) mouse-

tracking study found support for the default-intervention perspective, demonstrating that correct responders considered the heuristic response more than heuristic responders considered the correct response. However, Szaszi et al.'s (2017) think aloud study suggested that most correct-respondents began reasoning in line with the correct, not heuristic, strategy. The current study can be interpreted similarly to Szaszi et al.'s 2017 results, however, a key difference between these paradigms and the current study is the type of measurement used to examine reasoning strategies. These measurements may capture different cognitive phenomena to those captured through gaze measures.

To examine conflict on the CRT we compared heuristic-lure trials to correct-no lure trials. We found evidence for both explicit (confidence-based) and implicit (gaze-based) conflict during heuristic-lure trials. This is consistent with previous studies employing explicit confidence-based conflict measures (De Neys et al., 2013) and explicit 'feeling of error' based measure (Gangemi et al., 2015). However, our findings are inconsistent with Travers et al.'s (2016) examination of more implicit indicators of conflict. Travers et al. did not find evidence for a difference in reaction times between heuristic-lure and correct-no lure trials. Moreover, they found that participants moved their cursors to the heuristic option on lure items more quickly than they moved their cursors to the correct option on no lure items. Hence, they found no evidence that participants were slower to respond to heuristic-lure than correct-no lure items. In combination with the current findings, it is evident that the tools used to measure reasoning and conflict on the CRT can generate different results, possibly because they tap into distinct psychological phenomena.

In contrast to our hypotheses, we found no evidence that working memory dependence, strategy or conflict differed across test point. Some studies have demonstrated differences in working memory dependence across training on the CRT

(Purcell, Wastell, & Sweller, n.d.²³) while other studies have not (Purcell, Wastell, Sweller, et al., n.d.²⁴). In line with the literature on working memory and expertise (e.g., Ericsson & Kintsch, 1995), Purcell and colleagues proposed that working memory may play an important role in success on the CRT at intermediate points of numerical experience (Purcell, Wastell, & Sweller, n.d.²⁵). As such, we proposed that reflective strategies and conflict may be more common at intermediate points of training. Our study employed a cognitive load manipulation to examine differences in working memory dependence, however, the load condition did not show any effects. As such, there was no evidence that working memory dependence differed across test points. However, this null finding could be due to the load not placing sufficient demand on working memory resources or because this type of secondary task demands different cognitive resources to those required for the CRT. Therefore, alternative cognitive constraint paradigms, such as the two-response cognitive constraint method (Thompson et al., 2011), should be employed in future studies to confirm the relationship between training and working memory dependence.

Purcell, Wastell, Sweller, et al. (n.d.²⁶; Experiment 1) demonstrated an interaction between explicit conflict, test point and performance such that conflict became a stronger predictor of performance as training increased. The current study did not replicate this interaction, nor did it find evidence for an interaction between implicit conflict, test point and performance. Several differences between these studies may explain the inconsistent findings. For example, Purcell, Wastell, Sweller, et al. (n.d.²⁷; Experiment 1) study was conducted online and employed a two-response cognitive constraint paradigm rather than

^{23, 25} Paper 1 in the current thesis.

²⁴ Paper 2 in the current thesis.

^{26, 27} Paper 2 in the current thesis.

a cognitive load constraint. Therefore, future studies should examine whether working memory dependence, strategies, and conflict differ over training with alternative cognitive constraint paradigms to confirm if and how these factors are related.

The current study addressed an important gap in the literature on implicit indicators of strategy and conflict on the CRT. It incorporated eye tracking techniques to examine two new measures. The first, via fixations, was consistent with the hypothesis that participants would engage in greater information search by looking to the non-selected response options more often on heuristic lure trials, than they did on correct no lure trials. The second, via dwell, was consistent with the hypothesis that more time would be spent considering the alternatives for heuristic lure than correct no lure trials. These measures demonstrated strong convergent validity—an important consideration for the introduction of new measures. Interestingly, while the two gaze-based measures were strongly related, they were only weakly correlated with the confidence-based conflict measure. This weak association indicates that while participants exhibited signs of conflict on both explicit and implicit measures, these indicators may not be as strongly associated as previously thought. Future studies should examine the relationship between explicit and implicit conflict factors, and whether they make unique contributions to the engagement of working memory.

Theories of reasoning do not clearly distinguish between explicit and implicit conflict. The metacognitive dual process model asserts that the fluency of a process impacts the respondent's 'feeling of rightness' which can trigger the engagement of working memory (e.g., Thompson et al., 2011; Thompson & Morsanyi, 2012). The logical intuition dual process model proposes that competing processes may generate conflict which, in turn, can prompt the engagement of working memory dependent processes. While the metacognitive model highlights the role of subjective 'feelings of

rightness' and the logical intuition model highlights the role of underlying conflict, neither model clearly explains the interaction between these potentially distinct phenomena. To date, conflict studies have treated indicators of conflict that range from very explicit (e.g., think aloud protocols) to very implicit (e.g., autonomic arousal) as relatively interchangeable. However, the current study suggests that different types of conflict measures may reflect a range of cognitive processes. It is important to note that our study cannot speak to the relationship between conflict and working memory. However, future studies examining this key aspect of the hybrid models of reasoning may need to consider multiple indicators and constructs for cognitive conflict.

The CRT is a widely used task in reasoning and bias research. However, little is understood about the strategies employed or the role of cognitive conflict. The few studies that have investigated the nature of strategies and conflict on the CRT have produced inconsistent findings. The current study suggests that these inconsistencies could stem from the tools employed to measure the underlying cognitive processes. We demonstrated that explicit confidence-based measures and implicit gaze-based measures can be used to assess conflict but that they may reflect separate psychological phenomena.

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*, 1–43.
<https://doi.org/10.1080/13546783.2018.1507949>
- Bago, B., Raelison, 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., Phillips, P., Wade, C. N., & Quayle, J. D. (2006). Effects of Belief and Logic on Syllogistic Reasoning. *Experimental Psychology*. Göttingen, Germany :
<https://doi.org/10.1027/1618-3169.53.1.77>
- Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance*, 14(1), 12–23. <https://doi.org/10.1037/0096-1523.14.1.12>
- Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*. [Kidlington, Oxford, UK] : <https://doi.org/10.1016/j.tics.2004.10.003>
- 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*. Pully, Switzerland : <https://doi.org/10.3389/fpsyg.2019.00632>
- 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. Retrieved from <http://pps.sagepub.com>
- 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., 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., Moyens, E., & Ansteenwegen, D. V. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective and Behavioral Neuroscience*, 10(2), 208–216. <https://doi.org/10.3758/CABN.10.2.208>
- De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, 20(2), 269–273. <https://doi.org/10.3758/s13423-013-0384-5>
- De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science*. New York, NY :
<https://doi.org/10.1111/j.1467-9280.2008.02113.x>
- Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, 49(8), 709–724. <https://doi.org/10.1037/0003-066X.49.8.709>
- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2), 211–245. <https://doi.org/10.1037/0033-295X.102.2.211>
- 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>

- Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, 43(2), 226–236. <https://doi.org/10.3758/s13421-014-0461-7>
- Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Frey, D., & De Neys, W. (2017). Is Conflict Detection in Reasoning Domain General ? *Proceedings of the Annual Meeting of the Cognitive Science Society*, 39, 391–396. Retrieved from <https://pdfs.semanticscholar.org/aaec/4079bae9ba9ba75c6cf816874e5cc2b9a201.pdf>
- 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>
- Handley, S. J., & Trippas, D. (2015). Dual Processes and the Interplay between Knowledge and Structure: A New Parallel Processing Model. *The Psychology of*

- Learning and Motivation*. San Diego : <https://doi.org/10.1016/bs.plm.2014.09.002>
- Horstmann, N., Ahlgrimm, A., & Glöckner, A. (2009). How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes. *Judgment and Decision Making*, 4(5), 335–354.
- Johnson, E. D., Tubau, E., & De Neys, W. (2014). The unbearable burden of executive load on cognitive reflection: A validation of dual process theory The unbearable burden of executive load on cognitive reflection: A validation of dual process theory. *Proceedings of the Annual Meeting of the Cognitive Science Society*, (36), 36.
- Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, 164, 56–64.
<https://doi.org/10.1016/j.actpsy.2015.12.008>
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: 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: Cambridge University Press.
- 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*. [Chichester] . <https://doi.org/10.1002/bdm.1883>

- Purcell, Z. A., Wastell, C. A., & Sweller, N. (n.d.). Domain-Specific Experience and Dual-Process Thinking. [Paper 1 in the current thesis]
- Purcell, Z. A., Wastell, C. A., Sweller, N., & Howarth, S. (n.d.). No Pain, No Gain: Does Cognitive Conflict Predict Learning and Effortful Reasoning? [Paper 2 in the current thesis]
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*. [Washington, D.C.] : <https://doi.org/10.1037/0033-2909.124.3.372>
- Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, 135(6), 943–973. <https://doi.org/10.1037/a0017327>
- Sloman, S. (2014). Two systems of reasoning: An update. In *Dual-process theories of the social mind*. (pp. 69–79). New York, NY, US: The Guilford Press.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1), 3–22. <https://doi.org/10.1037/0033-2909.119.1.3>
- Stephen, D. G., Boncodd, 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>
- Stuppel, 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*. San Francisco, CA : <https://doi.org/10.1371/journal.pone.0186404>
- 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>

- Szaszi, Barnabas, 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. Retrieved from <http://simsrad.net.ocs.mq.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=127887902&site=ehost-live>
- 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>
- Thompson, V., & 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>
- 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>
- Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37(3), 498–505. <https://doi.org/10.3758/BF03192720>
- 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. (Boris M. Velichkovskii (Eds.), *Stratification in cognition and consciousness* (p. 203). Amsterdam, Netherlands: 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. [Exeter, U.K.] : <https://doi.org/10.1016/S1369->

8478(02)00013-X

Appendix A

Table A. *Examples of training items and feedback.*

Training Item	Incorrect Feedback
A pen and a notebook cost \$25 in total. The notebook costs \$5 more than the pen. How much does the pen cost?	<p>* Sorry, your answer was incorrect. Don't worry, we'll help you find the correct solution!</p> <p>The question stated: <i>A pen and a notebook cost \$25 in total. The notebook costs \$5 more than the pen. How much does the pen cost?</i></p> <p>It's often helpful to breakdown these problems into an algebraic format. Try and give it go!</p> <p>The question states "A pen and a notebook cost \$25 in total."</p> <p>How can we write this algebraically?</p> <p>Hint: let p represent "pen" and n represent "notebook"</p> <p>d) $p + n = 25$</p> <p>e) $p - n = 25$</p> <p>f) $p + 25 = n$</p>
If it takes 6 adults 6 minutes to blow up 6 balloons, how long would it take 12 adults to blow up 12 balloons?	<p>Sorry! Your response was incorrect.</p> <p>The original question stated: <i>If it takes 6 adults 6 minutes to blow up 6 balloons, how long would it take 12 adults to blow up 12 balloons?</i></p> <p>The correct answer is 6 minutes. Each person can blow up ONE balloon in 6 minutes, you have 12 adults now and over 6 minutes each of them blows up ONE balloon. There are 12 adults, so they can blow up 12 balloons in total.</p> <p>Try the next question.</p>
In the desert, there is an anthill. Every day, the anthill doubles in height. If it takes 4 days for the anthill to reach 2 meters tall, long would it take for the anthill to reach 1 meter tall?	<p>Sorry! Your response was incorrect.</p> <p>The question stated: <i>In the desert, there is an anthill. Every day, the anthill doubles in height. If it takes 4 days for the anthill to reach 2 meters tall, long would it take for the anthill to reach 1 meter tall?</i></p> <p>The correct answer is 3 days. If on the third day the anthill is 1 meter tall, and it doubles in height every day, the anthill will be 2 meters tall on the fourth day.</p> <p>See if you can get the next one!</p>

Notes. *Feedback for training items related to CRT item 1 include several algebraic steps. For brevity, they are not included but are available from the authors on request.

Appendix B

Table A. Means and standard deviations for performance, fixations and dwell (on the multiple-choice options) by test point, problem type and load condition.

Test Point	Problem Type	Load Condition	Performance <i>M (SD)</i>	Fixations <i>M (SD)</i>	Dwell (ms) <i>M (SD)</i>
1	Lure	Low	1.35 (1.09)	14.62 (8.16)	2956.68 (1771.49)
		High	1.44 (.92)	13.13 (5.47)	2857.59 (1234.44)
		Total	1.39 (1.00)	13.91 (6.96)	2909.75 (1521.23)
	No Lure	Low	2.95 (.22)	7.02 (3.11)	1380.17 (653.27)
		High	3.00 (.00)	6.04 (1.85)	1259.24 (335.53)
		Total	2.97 (.16)	6.55 (2.60)	1322.89 (524.04)
2	Lure	Low	2.30 (.80)	10.33 (4.70)	2318.32 (1049.28)
		High	2.61 (.70)	11.22 (8.03)	2607.69 (1946.44)
		Total	2.45 (.76)	10.75 (6.42)	2455.39 (1525.63)
	No Lure	Low	2.95 (.22)	4.95 (2.14)	1169.02 (430.24)
		High	3.00 (.00)	5.59 (2.22)	1329.93 (427.60)
		Total	2.97 (.16)	5.25 (2.18)	1245.24 (430.92)
3	Lure	Low	2.55 (.83)	9.67 (6.59)	2197.52 (1507.25)
		High	2.94 (.23)	9.41 (6.47)	2377.96 (1815.99)
		Total	2.73 (.64)	9.54 (6.44)	2282.99 (1640.16)
	No Lure	Low	2.95 (.22)	5.22 (2.43)	1177.28 (412.78)
		High	2.94 (.23)	5.28 (2.05)	1158.39 (382.89)
		Total	2.94 (.22)	5.25 (2.22)	1168.33 (393.63)
4	Lure	Low	2.45 (.60)	9.73 (5.74)	2149.75 (1114.56)
		High	2.55 (.62)	10.09 (8.34)	2453.07 (2249.62)
		Total	2.50 (.60)	9.90 (6.99)	2293.43 (1728.20)
	No Lure	Low	2.95 (.22)	4.53 (1.50)	970.20 (292.08)
		High	3.00 (.00)	5.70 (1.96)	1178.33 (465.10)
		Total	2.97 (.16)	5.08 (1.81)	1068.79 (392.79)

Appendix C

Table A. *Results for the linear mixed model used to examine the strategies (assessed via fixations) employed by correct and heuristic respondents across test points.*

	<i>F</i>	Df Source	Df Error	η_p^2	<i>p</i>
Item	.014	2	730.05	<.001	.986
Test Point	.003	3	837.29	<.001	>.999
Accuracy	.000	1	840.84	<.001	.987
AOI	150.530	3	1240.27	.267	<.001
Test Point * Accuracy	.001	3	819.60	<.001	>.999
Test Point * AOI	.964	9	1244.10	.007	.468
Accuracy* AOI	1.896	3	1244.30	.005	.128
Test Point * Accuracy*	1.255	9	1243.00	.009	.257
AOI					

Table B. *Results for the linear mixed model used to examine the strategies (assessed via dwell) employed by correct and heuristic respondents across test points.*

	<i>F</i>	Df Source	DF Error	η_p^2	<i>p</i>
Item	.005	2	701.74	<.001	.995
Test Point	.002	3	802.42	<.001	1.000
Accuracy	.002	1	806.40	<.001	.962
AOI	409.654	3	1221.00	.502	.000
Test Point * Accuracy	.001	3	785.17	<.001	1.000
Test Point * AOI	1.390	9	1224.29	.010	.187
Accuracy* AOI	1.813	3	1224.46	.004	.143
Test Point * Accuracy* AOI	1.395	9	1223.36	.010	.241

Note. *Df error* = 1750

Table C. *Pairwise comparisons for strategy via fixations.*

Comparison	Contrast Estimate	SE	Df source	F	η_p^2	<i>p</i>
1.00 - 2.00	0.304	.019	934.06	256.00	0.995	<.001
1.00 - 3.00	0.292	.017	1703.64	295.03	0.998	<.001
1.00 - 4.00	0.289	.017	1461.11	289.00	0.997	<.001
2.00 - 1.00	-0.304	.019	934.06	256.00	0.995	<.001
2.00 - 3.00	-0.012	.019	865.51	0.40	0.215	0.541
2.00 - 4.00	-0.015	.017	1726.95	0.78	0.517	0.364
3.00 - 1.00	-0.292	.017	1703.64	295.03	0.998	<.001
3.00 - 2.00	0.012	.019	865.51	0.40	0.215	0.541
3.00 - 4.00	-0.004	.019	861.33	0.04	0.029	0.853
4.00 - 1.00	-0.289	.017	1461.11	289.00	0.997	<.001
4.00 - 2.00	0.015	.017	1726.95	0.78	0.517	0.364
4.00 - 3.00	0.004	.019	861.33	0.04	0.029	0.853

Note. *Df error* = 1256.83. 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 D. *Pairwise comparisons for strategy via dwell.*

Comparison	Contrast Estimate	SE	Df source	F	η_p^2	<i>p</i>
1.00 - 2.00	.446	.017	914.533	688.29	0.998	<.001
1.00 - 3.00	.442	.015	1705.820	868.28	0.999	<.001
1.00 - 4.00	.428	.016	1478.661	715.56	0.999	<.001
2.00 - 1.00	-.446	.017	914.533	688.29	0.998	<.001
2.00 - 3.00	-.004	.017	849.432	0.06	0.037	.824
2.00 - 4.00	-.018	.015	1728.363	1.44	0.668	.233
3.00 - 1.00	-.442	.015	1705.820	868.28	0.999	<.001
3.00 - 2.00	.004	.017	849.432	0.06	0.037	.824

3.00 - 4.00	-.014	.017	846.107	0.68	0.317	.405
4.00 - 1.00	-.428	.016	1478.661	715.56	0.999	<.001
4.00 - 2.00	.018	.015	1728.363	1.44	0.668	.233
4.00 - 3.00	.014	.017	846.107	0.68	0.317	.405

Note. *Df error* = 1235.33. 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 E. Results for generalised linear mixed model used to examine the explicit conflict as measured via reverse-coded confidence ratings.

	<i>F</i>	Df source	η_p^2	<i>p</i>
Item	1.22	2	.004	.295
Test Point	3.16	3	.017	.024
Problem Type	107.32	1	.165	<.001
Test Point * Problem Type	3.20	3	.017	.023

Note. *Df error* = 543

Table F. Results for linear mixed model used to examine the implicit conflict as measured via the proportion of fixations.

	<i>F</i>	Df source	η_p^2	<i>p</i>
Item	.03	2	<.001	.968
Test Point	.03	3	<.001	.993
Problem Type	.09	1	<.001	.766
AOI	151.02	3	.173	<.001
Test Point * AOI	.78	9	.003	.634
Test Point * Problem Type	.03	3	<.001	.993
Problem Type * AOI	5.43	9	.007	.001
Test Point * Problem Type * AOI	1.16	9	.005	.320

Note. *Df error* = 2170

Table G. Results for linear mixed model used to examine the implicit conflict as

measured via the proportion of dwell.

	<i>F</i>	Df source	η_p^2	<i>p</i>
Item	.05	2	<.001	.953
Test Point	.01	3	<.001	.999
Problem Type	.02	1	<.001	.879
AOI	404.57	3	.359	<.001
Test Point * AOI	1.54	9	.006	.127
Test Point * Problem Type	.01	3	<.001	.999
Problem Type * AOI	6.07	9	.008	<.001
Test Point * Problem Type * AOI	.78	9	.003	.637

Note. *Df error* = 2170

Paper 4

Cognitive Conflict Predicts Working Memory Engagement on the Cognitive Reflection Test

Zoe A. Purcell¹, Stephanie Howarth², Colin A. Wastell¹, and Naomi Sweller¹

¹Department of Psychology, Macquarie University, Sydney, Australia

²Department of Cognitive Science, Macquarie University, Sydney, Australia

This paper has not been submitted for publication.

Abstract

Influential dual process theories of reasoning assert that there are two Types of processing: Type 1 which is fast and automatic, and Type 2 which is slow and requires working memory. The interaction between these Types is often examined using bias tasks which elicit tempting but incorrect responses. Over the past decade, research has revealed that reasoners are sensitive to conflicting logical principles even when they provide incorrect responses on bias tasks. As such, models of reasoning have proposed that conflict sensitivity may be involved in the engagement of Type 2, working memory dependent processes. The current study investigated whether conflict predicts working memory engagement, and whether implicit and explicit conflict make unique contributions. The study employed a two-response training paradigm. The confidence-based explicit conflict measure and gaze-based implicit conflict measure both predicted Type 2 processing. The findings support the proposition that conflict is associated with working memory engagement but also suggest that different methods of cognitive conflict detection may measure separate elements of conflict.

Cognitive Conflict Predicts Working Memory Engagement on the Cognitive Reflection Test

Reasoning and thinking research has generated many theories about why and how working memory is engaged. Several prominent models of reasoning come under the umbrella of dual process theory. Dual process theories make a distinction between Type 1, automatic and implicit thinking, and Type 2, deliberative and explicit thinking (e.g., Evans & Stanovich, 2013; Kahneman, 2011). One of the main distinguishing features of these processing Types is working memory engagement. Working memory is the cognitive “hardware” for holding and manipulating information in the short term (Baddeley, 1986; Hambrick & Engle, 2003). According to dual process theories, Type 2 processes are said to require working memory engagement, while Type 1 processes can operate with or without working memory (Evans & Stanovich, 2013; Thompson, 2013). A core line of enquiry for dual process theorists is how the two Types interact and, in particular, why, how, and for whom is Type 2 thinking activated.

The focus of reasoning research is shifting from the examination of the inter-individual factors, such as cognitive style and intelligence, associated with reflective thinking (e.g., Frederick, 2005; Toplak et al., 2011) to the intra-individual factors underlying the momentary engagement of effortful thinking (Bago & De Neys, 2017; Thompson et al., 2013). That is, the investigation into reasoning has become focussed not just on *who* is likely to engage in deeper thinking, but *when* and *how* that engagement is occurring. Contemporary dual process models of reasoning, known as hybrid dual process theories, have proposed that metacognitive factors such as conflict or “feelings of rightness” may be involved in the engagement of effortful thinking (e.g., De Neys, 2012; Pennycook, Fugelsang, & Koehler, 2015; Thompson & Morsanyi, 2012).

It seems plausible that an unexpected cue might elicit an explicit sense that something is not quite right, in other words, it may generate a lowered “feeling of

rightness”, or a heightened sense of conflict. Following these feelings of unease or incongruency, we may engage higher-level cognitive processes to, for example, examine our surroundings, gather more information, and act accordingly. However, it also seems plausible that the cue might have generated a form of cognitive uncertainty below the level of awareness (i.e. implicitly) and, subsequently, elicit information search and the engagement of deliberative processes prior to impacting our explicit sense of rightness or conflict. Previous studies have found support for the suggestion that conflict is involved in the engagement of deliberation, however, the potentially separate contributions of explicit and implicit conflict to that engagement have received limited empirical consideration.

The logical intuition model postulates that conflict may be involved in the engagement of Type 2 processing (De Neys, 2012, 2014). It suggests that multiple Type 1 processes may be triggered in response to a problem or task (see also Pennycook et al., 2015). Some triggered processes may have greater activation or salience than other triggered processes. The salience of a particular type, process or response can be increased through training (see Purcell, Wastell, & Sweller, n.d.²⁸; Bago & De Neys, 2019). If one of the processes’ salience is sufficiently greater than that of the competing processes, then this process will “win out” and generate the response. If, however, two or more processes have similar levels of activation, conflict is generated and potentially detected by the reasoner. According to the logical intuition model, that conflict—operating at an intuitive level—may be involved in the engagement of working memory dependent Type 2 processing.

The metacognitive dual process model integrated the dual process model of reasoning with the literature on metacognitive memory (Thompson, Prowse Turner, & Pennycook, 2011). Within this model, the ease and speed with which a solution comes to mind generates an affective metacognition. The easier and faster the solution is brought to

²⁸ Paper 1 in the current thesis.

mind, the more fluency the solution process has and, subsequently, the higher the reasoner's "feeling of rightness". This affective response, according to the metacognitive model of reasoning, is related to the engagement of Type 2 processing. The lower the "feeling of rightness" the higher the likelihood that working memory dependent, Type 2 processes will be engaged. Both the logical intuition and metacognitive dual process models propose some involvement of uncertainty in the engagement of Type 2 processes. However, there is limited research examining the relationship between conflict occurring at the implicit level and that occurring at more explicit levels.

A quintessential task used to investigate the engagement of Type 2 processing is the Cognitive Reflection Test (CRT; Frederick, 2005). The first item in the CRT, and the most widely studied, is the bat and ball problem. It 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 most common response is 10 cents; however, the correct answer is 5 cents. Reaching the correct solution has been interpreted as an indication that Type 2 processes have been engaged (Kahneman, 2011; Toplak et al., 2011). Early interpretations of CRT performance stipulated that correct responders first generate a heuristic incorrect response (10 cents) via Type 1 processes, and then override the incorrect response and generate the correct response (5 cents) via Type 2 processes (e.g., Kahneman, 2011; Kahneman & Frederick, 2005). However, research into the strategies employed on the CRT has produced mixed findings (Purcell, Howarth, Wastell, & Sweller, n.d.²⁹; Travers, Rolison, & Feeney, 2016). Moreover, there is evidence to suggest that some reasoners may be able to reach the correct solution under Type 1 conditions (e.g., Bago & De Neys, 2019).

There is evidence to suggest that reasoners are explicitly sensitive to conflict and fluency when providing erroneous responses on bias tasks, such as the CRT (e.g., De Neys,

²⁹ Paper 3 in the current thesis.

Rossi, & Houdé, 2013). Studies indicate that even when these tasks elicit incorrect responses, respondents may register their error. Conflict studies have compared confidence ratings and feelings of error on lure and no lure versions of the CRT. Lure versions reflect the original CRT items and are thought to cue both a tempting but incorrect response, as well as the correct response. In contrast, no lure items cue only a single response. The bat and ball problem described earlier is a lure item because it cues both the heuristic response (10 cents) and the correct response (5 cents). A no lure version of this item might read: A bat and a ball cost \$1.10 together. The bat costs \$1. How much does the ball cost? This question cues only the correct response: 10 cents. By comparing lure and no lure versions of the CRT items, conflict studies isolate the conflict induced by the lure task. De Neys et al. (2013) found that participants giving heuristic (i.e. incorrect but tempting) responses on the CRT reported lower confidence than they did when providing correct responses on no lure items (see also Hoover & Healy, 2019). Similarly, Gangemi, Bourgeois-Gironde and Mancini (2015) observed higher feelings of error for heuristic responding on lure items than correct responding on no lure items. These findings indicate that despite providing the incorrect, heuristic response, participants show explicit sensitivity to their error on the lure versions of the CRT.

Conflict studies have also examined more implicit measures of conflict. For example, Bago, Raelison and De Neys (2019) introduced a second-guess paradigm in which participants responded to the bat and ball problem and were then offered the chance to guess again. On their second attempt, responders showed a greater-than-chance likelihood of selecting a response that was smaller than their original, incorrect answer. That is, participants were more likely to choose a response that was closer to the correct answer. However, they rarely chose the correct option on their second attempt. This indicates that, despite not understanding how to determine the correct response, the respondents may have had some intuitive grasp that the answer is less than 10 cents.

Purcell, Howarth, et al. (n.d.³⁰) compared participants' eye-movements during correct responding on no lure items and incorrect responding on lure items. They found that the subjects examined the four multiple-choice options to a greater extent for lure items than no lure items. These studies suggest that reasoners giving heuristic responses to lure items demonstrate implicit sensitivity to the problems' mathematical principles.

Some studies have examined multiple measures of conflict simultaneously in paradigms employing the bat and ball problem (e.g., Frey, Johnson, & De Neys, 2018; Hoover & Healy, 2019). Multiple measures have been used to test their convergent validity rather than to consider the potentially distinct predictors of working memory engagement. In a comprehensive study, Hoover and Healy (2019) examined reasoners' recall and recognition for the critical "more than" phrase in the bat and ball problem, as well as their post-decision confidence and their opinion on whether other reasoners could answer the lure item. These measures showed convergent effects regarding the differences between correct and incorrect reasoning and lure and no lure items, however, the authors point out that, despite shared variability, different measurements may index separate error signals. Purcell, Howarth, et al. (n.d.³¹) demonstrated convergent effects for explicit and implicit conflict measures employing confidence ratings, eye fixations and dwell time, when comparing lure and no lure items. Moreover, they found that the three measures were significantly associated, but with only weak correlations between the explicit confidence factor and the two implicit gaze-based factors. In line with Hoover and Healy (2019), Purcell et al. suggested that the measurements may index separate conflict-based phenomena. Although there is growing evidence to suggest distinctions between the different conflict indicators, multiple measures of conflict have not been tested as potentially separate predictors of working memory engagement.

^{30, 31} Paper 3 in the current thesis.

To test the relationship between conflict and working memory engagement, researchers have employed various cognitive constraint techniques. Cognitive constraint manipulations help the experimenter to classify a reasoner's thinking Type by reducing the reasoner's temporary capacity to engage Type 2 processes. Type 2 processing is thought to be more time-consuming and more cognitively demanding than Type 1 processing (e.g., Kahneman, 2011). Therefore, cognitive constraint studies aim to manipulate the conditions under which the participants are reasoning, for example using time-restrictions or cognitive load tasks. One of the most well-known cognitive constraint techniques is known as the two-response paradigm. Participants are given two opportunities to answer the problems, once while instructed to give the first response that comes to mind (sometimes under a time limit), and again without restrictions (Thompson et al., 2011, 2013). Those able to reach the correct solution under the constraint (i.e. at Response 1) are thought to have done so via Type 1 processing, whereas those who reached the correct response only under conditions without the constraint (i.e. at Response 2) are thought to have used Type 2 processing.

Studies examining performance on the CRT in conjunction with a two-response paradigm found that some of the respondents who provide correct responses were able to do so under Type 1 conditions. For example, Thompson et al. (2013) found that some participants were able to successfully complete the CRT under timing restrictions. Others demonstrated that some participants were even able to reach correct solutions under both time-limit and an additional cognitive load (Bago & De Neys, 2019). However, the number of participants who can complete the CRT under Type 1 conditions is usually low, as is the number of participants who change their responses from Response 1 (constrained) to Response 2 (unconstrained).

Further, the chances of a participant exhibiting both Type 1 processing and Type 2 processing during the one study is also low (e.g., Bago & De Neys, 2017; see Purcell,

Wastell, Sweller, & Howarth, n.d.³²), because reasoning research employing bias tasks like the CRT typically employs single-test paradigms. That is, the participant completes the task in one sitting and their reasoning is classified as having been Type 1 or Type 2 for each problem completed. This typically leads to low instances of Type 2 engagement and high stability indices on bias tasks (i.e., each participant is likely to exhibit only Type 1 or Type 2 processing; e.g., Bago & De Neys, 2017; Pennycook & Thompson, 2012; Thompson & Johnson, 2014; Thompson et al., 2011). As previous studies have noted, this high stability reduces the instances of within-subject comparisons of Type 1 and Type 2 processing (e.g., Bago & De Neys, 2019; Purcell, Wastell, Sweller, et al., n.d.³³). This stability is not overly concerning for examining the relationship between thinking Type and individual difference factors. However, it is problematic for the examination of intra-individual factors, because any comparisons between Type 1 and Type 2 processing is consequently largely between-subject (see Bago & De Neys, 2017). When examining the intra-individual factors like the cognitive mechanisms underlying the momentary engagement of Type 2, working memory dependent processes, it is important to compare thinking Types as a within-subject design where possible. To address this issue, Purcell, Wastell, Sweller, et al. (n.d.³⁴) incorporated a mathematical training paradigm with a two-response manipulation. By increasing the participants' domain-specific experience through training during the study, they increased the likelihood that participants would demonstrate both Type 1 and Type 2 thinking.

The logical intuition model asserts that some logical intuitions are developed through practice. Purcell, Wastell, and Sweller (n.d.³⁵) incorporated the literature on expertise and working memory (e.g., Ericsson & Kintsch, 1995; Hambrick & Oswald,

^{32, 33, 34} Paper 2 in the current thesis.

^{35, 36, 40} Paper 2 in the current thesis.

2005), and empirically tested this assumption using the CRT. They found that, as domain-specific experience increased with training, the reasoner's dependence on working memory initially increased and then declined. This pattern of working memory dependence was interpreted as an indication that working memory was important for intermediate levels of domain-specific knowledge, that is, when the reasoner had sufficient knowledge of the relevant logical principles but insufficient practice for automation. Additionally, they suggested that the salience of the relevant logical principles may have increased with training, and that as the salience of the logical principle surpassed that of the heuristic principle, greater conflict may have been experienced, which, in turn corresponded with the instances of higher working memory dependence. However, later training studies using different versions of CRT-like problems at each test did not find evidence of automation (Purcell, Howarth, et al., n.d.³⁶; Purcell, Wastell, Sweller, et al., n.d.³⁷).

Purcell, Wastell, Sweller, et al. (n.d.³⁸) did not find evidence for systematic differences in working memory dependence with increasing mathematical training. However, the paradigms did elicit high within-subject variability in thinking Types (also known as a low stability index), which facilitated the within-subject examination of intra-individual factors. The Purcell et al. (Purcell, Howarth, et al., n.d.³⁹; Purcell, Wastell, Sweller, et al., n.d.⁴⁰) training paradigm elicited lower stability indices to yield more rigorous examination of mechanisms, like cognitive conflict, proposed to impact the engagement of Type 2 thinking. Therefore, to examine whether conflict is involved in the engagement of Type 2 thinking, the current paradigm incorporated a two-response technique with a training manipulation to increase the likelihood of capturing instances of Type 1 and Type 2 thinking within the same individual.

^{37, 38, 39} Paper 3 in the current thesis.

Three research questions were generated to investigate the relationship between conflict and Type 2 processing: 1) Do explicit and implicit conflict occur during responding on the CRT and does explicit or implicit conflict change with practice? 2) Does explicit or implicit conflict predict working memory engagement, and does this relationship change with practice? And, 3) are explicit and implicit conflict related to each other, and do they make unique contributions to the engagement of working memory? We explored these questions using a CRT training program with four test points (T1, T2, T3, T4), and two problem types (lure and no lure). A two-response cognitive constraint paradigm was employed at each test point such that respondents answered each problem twice – once under a time-limit (Response 1) and again with no time-limit (Response 2). Additionally, it included two measures of conflict, an explicit confidence-based measure and an implicit eye-tracking based measure (Purcell, Howarth, et al., n.d.⁴¹). There were four sets of hypotheses for the study, corresponding to the behavioural aspects and the three research questions. These hypotheses are explained in more detail below.

Behavioural. In line with Purcell, Wastell, Sweller, et al. (n.d.⁴²) we expected that performance would increase with test point and that performance would be higher at Response 2 than Response 1.

Conflict. The second set of hypotheses examined the relationship between test point and trial type on explicit and implicit conflict. As in previous studies, explicit conflict was measured through reverse coded confidence ratings (e.g., De Neys, Cromheeke, & Osman, 2011; Purcell, Wastell, Sweller, et al., n.d.⁴³). Implicit conflict was examined in two steps. First, an implicit conflict factor was created by taking the inverse proportion of fixations that occurred on the selected answer. For example, if 20% of the fixations occurred on the

⁴¹ Paper 3 in the current thesis.

⁴² Paper 2 in the current thesis.

^{43, 46} Paper 2 in the current thesis.

option that was finally selected, the implicit conflict factor would be 80% (Purcell, Howarth, et al., n.d.⁴⁴). Second, we examined implicit conflict by the spread of fixations (moments when gaze is stationary) across the multiple-choice alternatives (Purcell, Howarth, et al., n.d.⁴⁵). The greater the spread of fixations, the greater the implicit conflict. We expected that conflict would be higher for lure problems on which participants gave heuristic responses than no lure problems on which participants gave correct responses. We expected this relationship to emerge for both explicit and implicit conflict.

Conflict and working memory engagement. The third set of hypotheses concerned the relationship between conflict and working memory engagement. As in previous studies, answer change between Response 1 and Response 2 was interpreted as an indication of working memory engagement (e.g., Bago & De Neys, 2017; Purcell, Wastell, Sweller, et al., n.d.⁴⁶). As for the hypotheses above, an explicit conflict factor was generated by reverse coding confidence ratings and an implicit conflict was generated by taking the inverse proportion of fixations that occurred on the selected answer. For both explicit and implicit conflict, we expected that higher conflict would predict a greater likelihood of working memory engagement (i.e. answer change from Response 1 to Response 2).

The relationship between explicit conflict, implicit conflict and working memory engagement. Additionally, we examined whether implicit and explicit conflict factors were related, and whether they made unique contributions to the engagement of working memory. We would like to stress that the classification of conflict as implicit or explicit is a simplification and these phenomena are likely to occur on a continuum. However, we believe it is important to consider that different indicators of conflict may have different

^{44, 45} Paper 3 in the current thesis.

implications for reasoning (Purcell, Howarth, et al., n.d.⁴⁷). Confidence ratings have been used as indicators of implicit cognitive conflict, yet they require an explicit awareness of a cognitive conflict in order for the reasoner to report that conflict. Eye-tracking offers an alternative measurement for implicit conflict that does not rely on explicit awareness. While explicit measures of conflict have been found to correlate with working memory engagement, it is not clear if this explicit conflict is simply an indicator of an implicit process, or whether the explicit conflict is, in fact, a unique contributor to the engagement of working memory. The examination of the relationship between the explicit and implicit measures of conflict was an exploratory investigation and so no specific hypotheses were generated.

Method

Participants & Design

A 4 (test point: T1, T2, T3, T4) x 2 (response type: Response 1 and Response 2) x 2 (problem type: lure and no lure) within-subjects design was used. Participants were 38 undergraduate psychology students at Macquarie University (Sydney, Australia) awarded course credit for participation. All participants had normal vision. Participants were 24 females and 14 males with ages ranging from 18 to 41 ($M=21.24$, $SD=4.65$).

Apparatus

Participants were tested individually with the experimenter present⁴⁸. The reasoning task was completed on a 24.5-inch LCD monitor (BenQ XL2540, refresh rate 240 Hz, natural resolution 1920 x 1080) and right-eye movements were recorded with a desk mounted eye-tracker sampling at a rate of 1000 Hz (EyeLink 1000; SR Research Ltd.,

⁴⁷ Paper 3 in the current thesis.

⁴⁸ A working memory capacity test was conducted in the same session for use in a different study, the order of the working memory test and the reasoning task were counterbalanced. The working memory test took approximately ten minutes to complete. [This study is not presented in the current thesis.]

Osgoode, Ontario, Canada). Participants used a chinrest to maintain a viewing distance of 800 mm and reduce head movements. Prior to commencing the reasoning training task, nine-point eye-tracking calibration and validation were conducted. 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. The reasoning training task was run on Experiment Builder 1.10.165 (SR-Research) and the data were extracted using Eyelink Data Viewer (SR-Research).

Eye-tracking measures were taken throughout the reasoning task. For coding, Areas of Interest (AOIs) were allocated to each of the multiple-choice alternatives. The questions were presented with four multiple choice response options that participants could choose from (see Figure 1D and 1F). These were labelled ‘Correct’ for the correct answer, and ‘Incorrect-1’, ‘Incorrect-2’ and ‘Incorrect-3’ for the other responses. For lure items, ‘Incorrect-1’ represented the incorrect but tempting (i.e. heuristic) response. In contrast, for no lure items, there was no important distinction between the incorrect responses. Importantly, the AOIs were dynamic such that, as the options were randomly allocated to corners of the screen, the AOI reflected the response option not the screen location. The number of fixations and dwell (sum of the duration of the fixations) was recorded for each AOI.

Measures

Participants completed a reasoning training task that has been used to improve performance on CRT-like problems in previous studies (Purcell, Howarth, et al., n.d.⁴⁹; Purcell, Wastell, Sweller, et al., n.d.⁵⁰). The task included lure and no lure problems. All problems were based on one of the three original CRT questions (Frederick, 2005). The lure problems mirrored the structure of the original questions; however, the content and

⁴⁹ Paper 3 in the current thesis.

⁵⁰ Paper 2 in the current thesis.

quantities were changed to prevent rote learning or recognition effects throughout training. Each lure item had a corresponding no lure item which had a similar structure but with no tempting incorrect response. Items reflected the word lengths of the original CRT item (+/- 1 word). An example of a conflict 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-conflict problem was: “A phone and a wallet cost \$1000 in total. The phone costs \$800. How much does the wallet cost?” (Answer: \$200). These items and training protocols have been used to improve participants’ performance in previous studies (Purcell, Howarth, et al., n.d.⁵¹; Purcell, Wastell, Sweller, et al., n.d.⁵²).

Procedure

Participants gave consent prior to completing the study. At the beginning of the reasoning task, participants were given written and verbal instructions about the general procedure including a brief explanation of the eye-tracking equipment. The calibration was then conducted, and the reasoning task began. Two three-minute breaks were included throughout the task. During the breaks, participants were advised to sit back from the chin rest and close their eyes to hydrate them and decrease blinking, and to reduce potential fatigue effects.

Participants were presented with seven blocks of six maths problems (three lure and three no lure items), in the order: T1, training-block 1, T2, training-block 2, T3, training-block 3, and T4. The order of the blocks was counterbalanced and the order of items within each block was randomised. A two-response paradigm was employed in the test blocks (e.g., Thompson et al., 2011). Each maths problem was presented twice, once with a time limit imposed (Response 1) and again with no time-limit (Response 2). Each trial began

⁵¹ Paper 3 in the current thesis.

⁵² Paper 2 in the current thesis.

with a fixation cross, presented for 3000ms (Figure 1A). The maths problem was then presented, one sentence at a time; sentences were presented cumulatively for 3000ms each (Figure 1B and 1C)). The final sentence was presented with the multiple-choice alternatives and remained on the screen until a response was made or it timed-out after 5000ms (Figure1D).

<p>(A)</p> <p>+</p>	<p>(B)</p> <p>A bag and a badge cost \$12.10 in total.</p>	<p>(C)</p> <p>A bag and a badge cost \$12.10 in total. The bag costs \$2.00 more than the badge.</p>	<p>(D)</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 25%; text-align: center;">5.05</td> <td style="width: 25%;"></td> <td style="width: 25%; text-align: center;">10.10</td> <td style="width: 25%;"></td> </tr> <tr> <td colspan="4" style="text-align: center;"> <p>A bag and a badge cost \$12.10 in total. The bag costs \$2.00 more than the badge.</p> <p>How much does the badge cost?</p> </td> </tr> <tr> <td style="text-align: center;">7.05</td> <td></td> <td style="text-align: center;">11.10</td> <td></td> </tr> </table>	5.05		10.10		<p>A bag and a badge cost \$12.10 in total. The bag costs \$2.00 more than the badge.</p> <p>How much does the badge cost?</p>				7.05		11.10	
5.05		10.10													
<p>A bag and a badge cost \$12.10 in total. The bag costs \$2.00 more than the badge.</p> <p>How much does the badge cost?</p>															
7.05		11.10													
<p>(E)</p> <p>Did you give the first response that came to mind?</p> <p>a) Yes b) No</p> <p>How confident are you in this response? (0=not at all confident, 100=absolutely confident)</p> <p>0----- -----100</p>	<p>(F)</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 25%; text-align: center;">11.10</td> <td style="width: 25%;"></td> <td style="width: 25%; text-align: center;">10.10</td> <td style="width: 25%;"></td> </tr> <tr> <td colspan="4" style="text-align: center;"> <p>A bag and a badge cost \$12.10 in total. The bag costs \$2.00 more than the badge.</p> <p>How much does the badge cost?</p> </td> </tr> <tr> <td style="text-align: center;">7.05</td> <td></td> <td style="text-align: center;">5.05</td> <td></td> </tr> </table>	11.10		10.10		<p>A bag and a badge cost \$12.10 in total. The bag costs \$2.00 more than the badge.</p> <p>How much does the badge cost?</p>				7.05		5.05		<p>(G)</p> <p>How confident are you in this response? (0=not at all confident, 100=absolutely confident)</p> <p>0----- -----100</p>	
11.10		10.10													
<p>A bag and a badge cost \$12.10 in total. The bag costs \$2.00 more than the badge.</p> <p>How much does the badge cost?</p>															
7.05		5.05													

Figure 1. *Sample trial presentation. Screens were displayed in alphabetical order. Note: Some proportions are changed in this figure for readability.*

Immediately after providing an answer (Response 1), participants who selected an answer before the question timed-out were asked if they gave the first response that came to mind and how confident they were in that response (Figure 1E). Participants who did not respond before the question timed-out 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. Participants who did not give a response before a question timed out were not presented with the option of reporting their confidence. The maths problem was then presented again to all participants, this time the full question and the answer options were presented at once (Figure 1F) and displayed until an answer was selected (Response 2). After providing Response 2, participants were asked how confident they were in this response (Figure 1G).

Test blocks were separated by training blocks. Training blocks differed from test blocks in that the participants were presented each problem once, without a timing constraint, and feedback was provided. Participants were advised whether their response was correct or incorrect, presented with the full question and answer, then given a brief explanation as to why that was the correct solution. Examples of feedback and guidance are provided in Appendix Table A. These items and training protocols have been used to improve participants' performance in previous studies (Purcell, Howarth, et al., n.d.⁵³; Purcell, Wastell, Sweller, et al., n.d.⁵⁴).

Results

Behavioural Results

Descriptive analyses for performance are reported in Table 1 including averages across the key factors: test point (T1, T2, T3, T4), response type (response 1, response 2),

⁵³ Paper 3 in the current thesis.

⁵⁴ Paper 2 in the current thesis.

and problem type (lure, no lure). The pattern of performance on the lure items by test point and response type is formally analysed below.

Table 1. *Means and standard deviations of performance by test point, response type (Response 1 and Response 2), and problem type (lure and no lure). Scores could range from 0-3, N=38.*

		Lure Items	No Lure Items	Average
		<i>M (SD)</i>	<i>M (SD)</i>	<i>M (SD)</i>
Test Point 1	Response 1	.50 (.80)	2.34 (.75)	1.42 (1.20)
	Response 2	1.34 (1.10)	2.82 (.39)	2.08 (1.11)
	Average	.92 (1.05)	2.58 (.64)	1.75 (1.20)
Test Point 2	Response 1	1.13 (.91)	2.45 (.69)	1.79 (1.04)
	Response 2	2.24 (.94)	2.95 (.23)	2.59 (.77)
	Average	1.68 (1.07)	2.70 (.57)	2.19 (1.00)
Test Point 3	Response 1	1.74 (.92)	2.61 (.60)	2.17 (.89)
	Response 2	2.37 (.82)	2.95 (.23)	2.66 (.66)
	Average	2.05 (.92)	2.78 (.48)	2.41 (.82)
Test Point 4	Response 1	1.66 (.97)	2.71 (.61)	2.18 (.96)
	Response 2	2.24 (.82)	2.89 (.31)	2.57 (.70)
	Average	1.95 (.94)	2.80 (.49)	2.37 (.86)
Average	Response 1	1.26 (1.02)	2.53 (.67)	1.89 (1.07)
	Response 2	2.05 (1.01)	2.90 (.30)	2.47 (.86)
	Average	1.65 (1.09)	2.71 (.55)	2.18 (1.01)

To formally examine the behavioural results, a repeated measures ANOVA was used to examine the effects of response type (Response 1, Response 2) and test point (T1,

T2, T3, T4) on performance on lure items. The ANOVA showed significant main effects of response type, $F(1,37)=94.07$, $p<.001$, $\eta_p^2=.718$, and test point $F(3,111)=31.92$, $p<.001$, $\eta_p^2=.463$. Performance was lower at Response 1 ($M=1.26$, $SE=.10$) than at Response 2 ($M=2.05$, $SE=.12$). Performance was higher at T2 ($M=1.68$, $SE=.14$) than at T1 ($M=.92$, $SE=.13$), $F(1,35)=27.69$, $p<.001$, $\eta_p^2=.442$, and higher at T3 ($M=2.05$, $SE=.12$) than at T2, $F(1,35)=10.61$, $p=.002$, $\eta_p^2=.233$, but not significantly different between T3 and T4 ($M=1.95$, $SE=.13$), $F(1,35)=1.00$, $p=.324$, $\eta_p^2=.028$. This indicates that performance improved from T1 to T2 and T2 to T3, but remained stable from T3 to T4.

Conflict

Explicit Conflict

Confidence ratings for Response 1 were reverse coded to create explicit conflict scores. Participants could not give confidence ratings on items that had ‘timed-out’. Therefore, any items on which the question timed-out before a response was made was excluded for this analysis. As in previous studies of conflict (Purcell, Howarth, et al., n.d.⁵⁵), we compared heuristic-lure (HL) trials to correct-no lure (CNL) trials, that is, we compared participants’ conflict when they had provided the tempting but incorrect response on lure items, to participants’ conflict when they had provided the correct response on no lure items. A linear mixed model was used to analyse the effects of test point (T1, T2, T3, T4) and trial type (CNL, HL) on explicit conflict scores. Trial type was nested within item⁵⁶, item within test point, and test point within participant. There was a significant two-way interaction between trial type and test point (see Table 2), suggesting that the differences in conflict scores for HL and CL trials varied across test point.

Table 2. *The linear mixed model used to examine the effects of item, test point and*

⁵⁵ Paper 3 in the current thesis.

⁵⁶ Item is used here and in the following analyses to refer to the structure of the problem reflecting items 1 to 3 in the original CRT.

trial type on explicit conflict.

Source	Df _{Source}	Df _{Error}	<i>F</i>	η_p^2	<i>p</i>
Item	2	374.11	6.84	.035	.001
Test Point	3	391.57	3.05	.023	.029
Trial Type	1	254.24	142.62	.359	<.001
Test Point * Trial Type	3	273.92	5.69	.059	.001

To examine the interaction more closely, trial types were compared within each test point. The difference between conflict on HL and CNL trials was significant at all test points, such that conflict was greater for HL trials than CL trials. At T1, for HL trials $M=32.28$, $SE=3.45$, for CNL trials $M=19.54$, $SE=2.90$, $F(1,86.33)=14.68$, $p<.001$, $\eta_p^2=.145$. At T2, for HL trials $M=47.05$, $SE=4.29$, for CNL trials $M=17.68$, $SE=2.29$, $F(1,55.56)=45.75$, $p<.001$, $\eta_p^2=.452$. At T3, for HL trials $M=46.02$, $SE=4.70$, for CNL trials $M=16.73$, $SE=2.11$, $F(1,48.25)=40.23$, $p<.001$, $\eta_p^2=.455$. At T4, for HL trials $M=35.09$, $SE=3.38$, for CNL trials $M=13.92$, $SE=2.01$, $F(1,78.99)=38.84$, $p<.001$, $\eta_p^2=.330$. Although the differences between the HL and CNL conflict scores were significant and in the same direction at each test point, the effect sizes suggest that trial type was a stronger predictor of conflict at the later test points (T2, T3, T4) than at the initial test point (T1).

Implicit Conflict

We conducted a two-step examination of the pattern of implicit conflict across test point (T1, T2, T3, T4) and trial type (CNL, HL). First, we used a single factor measure of implicit conflict, and second, we examined the spread of fixations across the four multiple choice options (Purcell, Howarth, et al., n.d.⁵⁷). A single factor for implicit conflict was

⁵⁷ Paper 3 in the current thesis.

calculated as the inverse of the proportion of fixations that occurred during Response 1. A linear mixed model was used with trial type nested within item, item nested within test point, and test point within participant. The model included item, test point and trial type as predictors, and the single factor implicit conflict score was the dependent variable. The model did not show any significant effects (see Table 3). As such there was no evidence for differences between implicit conflict on HL compared to CNL trials at any test point.

Table 3. *The linear mixed model used to examine the effects of item, test point and trial type on implicit conflict.*

Source	Df _{Source}	Df _{Error}	<i>F</i>	η_p^2	<i>p</i>
Item	2	461.53	.47	.002	.625
Test Point	3	371.75	.19	.002	.902
Trial Type	1	298.79	.30	.001	.586
Test Point * Trial Type	3	322.13	.53	.005	.660

To examine the spread of the proportion of fixations that occurred in each AOI (corresponding to the four multiple-choice alternatives) for each trial type (HL, CNL), the mean proportions of fixations were examined across test point, trial type, and AOI (see Figure 3). The differences observed in Figure 3 were formally tested using a linear mixed model. AOI was nested within trial type, trial type within item, item within test point, and test point within participant. The three-way interaction between test point, trial type, and AOI was not significant, $F(9,1605.02)=.516$, $p=.864$, $\eta_p^2=.003$ (outcomes for the full model are provided in Appendix Table B). However, the interaction between trial type and AOI was significant, $F(9,1602.23)=52.69$, $p<.001$, $\eta_p^2=.090$. Therefore, simple effects of AOI were examined for CNL and HL trials separately.

For CNL items, the simple effect of AOI was significant, $F(3,1227.68)=215.89$, $p<.001$, $\eta_p^2=.345$. For these no lure items, the AOIs for ‘Incorrect-1’, ‘Incorrect-2’ and

‘Incorrect-3’ had no theoretical difference; they represent multiple-choice options that are neither correct nor heuristic. Therefore, these AOIs were combined for the follow up test.

For CNL trials, the proportion of fixations on the ‘Correct’ AOI was greater than the proportions of fixations on the incorrect AOIs combined (mean difference=.28),

$F(1,1281.88)=619.91, p<.001, \eta_p^2=.326$.

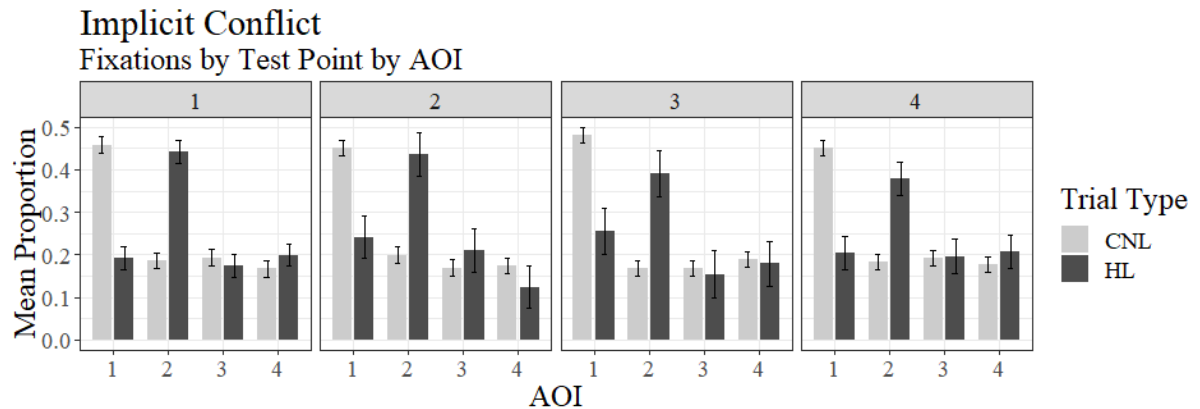


Figure 3. Mean proportions of fixations by test point, trial type and AOI. AOIs:

1=Correct, 2=Incorrect-1, 3=Incorrect-2, 4=Incorrect-3. Error bars reflect +/- 1 SE.

For HL trials, the simple effect of AOI was significant, $F(3,275.35)=39.42, p<.001, \eta_p^2=.300$. For these lure items, the AOIs for ‘Incorrect-2’ and ‘Incorrect-3’ had no theoretical difference; they represent the multiple-choice options that are neither correct nor heuristic. Therefore, these two AOIs were combined for the follow up contrast tests. Three follow up contrasts were run which compared the ‘Correct’ AOI to the ‘Incorrect-1’ AOI, the ‘Correct’ AOI to the ‘Incorrect-2’ and ‘Incorrect-3’ AOIs combined, and the ‘Incorrect-1’ AOI to the ‘the ‘Incorrect-2’ and ‘Incorrect-3’ AOIs combined. These results were compared to a Bonferroni adjusted alpha of .017. The proportion of fixations on the ‘Correct’ AOI was significantly lower than the proportion of fixations on the ‘Incorrect-1’ AOI (mean difference=.21), $F(1, 185.28)=47.04, p<.001, \eta_p^2=.202$. The proportion of fixations on the ‘Incorrect-1’ AOI was significantly higher than the proportion of fixations on the ‘Incorrect-2’ and ‘Incorrect-3’ AOIs (mean

difference=.24), $F(1, 343.18)=108.38$, $p<.001$, $\eta_p^2=.240$. Notably, for HL trials, the proportion of fixations on the ‘Correct’ AOI was not significantly different to the proportion of fixations on the ‘Incorrect-2’ and ‘Incorrect-3’ AOIs (mean difference=.053), $F(1, 272.39)=5.61$, $p=.019$, $\eta_p^2=.020$. This non-significant result suggested that participants did not look to the correct response on HL trials to a greater extent than the other incorrect responses, however, given the conservative Bonferroni adjustment, this interpretation was tentative.

These results indicate that the proportion of fixations was greatest for the correct response on CNL trials and the heuristic response for HL trials. That is, the proportion of fixations was greatest for the selected response on both CNL and HL trials. This consistency in results between CNL and HL trials suggests that implicit conflict was not occurring to a greater extent on HL trials relative to CNL trials. However, this conclusion rests on a conservative Bonferroni adjustment and may reflect a power issue rather than a null finding. As for the results reported above, there was no evidence that the pattern of implicit conflict changed across test point.

Conflict & Working Memory Engagement

To examine the relationship between conflict and working memory engagement, a new ‘working memory engagement’ variable was generated. First, items were coded to reflect change categories. The answers on lure items were coded with a two-digit number to reflect participants’ answers (correct or incorrect) at Response 1 and Response 2. Trials on which participants gave a correct answer at Response 1 and Response 2 were coded ‘11’, trials with both incorrect answers were coded ‘00’. Trials on which participants gave a correct answer at Response 1 and an incorrect answer at Response 2 were coded ‘10’ and, finally, trials on which participants gave an incorrect answer at Response 1 and a correct answer at Response 2 were coded ‘01’. Table 4 describes the general patterns of

answer change across the test points.

Table 4. *Number of trials in each change category by test point.*

Test Point	Change Category				Total
	00	01	10	11	
1	58	34	4	15	111
2	30	40	0	43	113
3	21	27	4	62	114
4	29	24	1	62	116
Total	138	125	9	182	454

Note: *Time-out trials are included as ‘incorrect’ items (i.e. coded ‘0’) for this analysis.*

Second, the change codes were transformed into a binary ‘working memory engagement’ variable for use as the dependent variable in the following analyses. As in previous studies, answer changes were interpreted as an indication that working memory had been engaged (Bago & De Neys, 2017; Purcell, Wastell, Sweller, et al., n.d.⁵⁸). Therefore, change codes (01 and 10) were pooled to reflect working memory engagement and no-change codes (11 and 00) were pooled to reflect no working memory engagement. Working memory engagement was indicated on 29.52% of trials.

Explicit Conflict and Working Memory Engagement

As above, explicit conflict was calculated by reverse scoring confidence ratings. Participants could not give confidence ratings on items that had ‘timed-out’. Therefore,

⁵⁸ Paper 2 in the current thesis.

‘time-out’ items were excluded for this analysis. Working memory engagement by explicit conflict scores are reported in Figure 4. A binary logistic generalised linear mixed model was used with conflict nested within item, item nested within test point, and test point within participant. The dependent variable was working memory engagement and the predictors were item, test point and explicit conflict. The model revealed significant effects of item, $F(2, 364)=9.94$, $p<.001$, $\eta_p^2=.052$, and test point, $F(3, 364)=4.60$, $p=.004$, $\eta_p^2=.037$, on working memory engagement. Follow up tests revealed that working memory engagement was more likely for items 1 and 3 of the CRT than item 2 (see Appendix Table C), and that working memory engagement was less likely as training increased, and (see Appendix Table D). Additionally, explicit conflict had a significant effect on working memory engagement, $F(1, 364)=20.55$, $p<.001$, $OR=1.021[CI=1.012, 1.031]$. This indicated that greater explicit conflict was associated with a higher likelihood that the participant engaged working memory.

Implicit Conflict and Working Memory Engagement

As in earlier analyses, a single-factor implicit conflict score was calculated as the inverse of the proportion of fixations on the selected response (Purcell, Howarth, et al., n.d.⁵⁹). Working memory engagement by implicit conflict scores (as a percentage) is reported in Figure 4. A binary logistic generalised linear mixed model was conducted with working memory engagement as the dependent variable and predictors: item, test point and implicit conflict. The model revealed significant effects of item, $F(2, 364)=11.38$, $p<.001$, $\eta_p^2=.059$, and test point, $F(3, 364)=6.05$, $p=.001$, $\eta_p^2=.047$, on working memory engagement. As above, working memory engagement was more likely for items 1 and 3 than item 2 (see Appendix Table E), and decreasingly likely as training

⁵⁹ Paper 3 in the current thesis.

increased (see Appendix Table F). Additionally, implicit conflict had a significant effect on working memory engagement, $F(1, 364)=4.73$, $p=.030$, $\eta_p^2=.013$, $OR=4.216$ [$CI=1.148, 15.487$]. This indicated that greater implicit conflict was associated with a higher likelihood of working memory engagement.

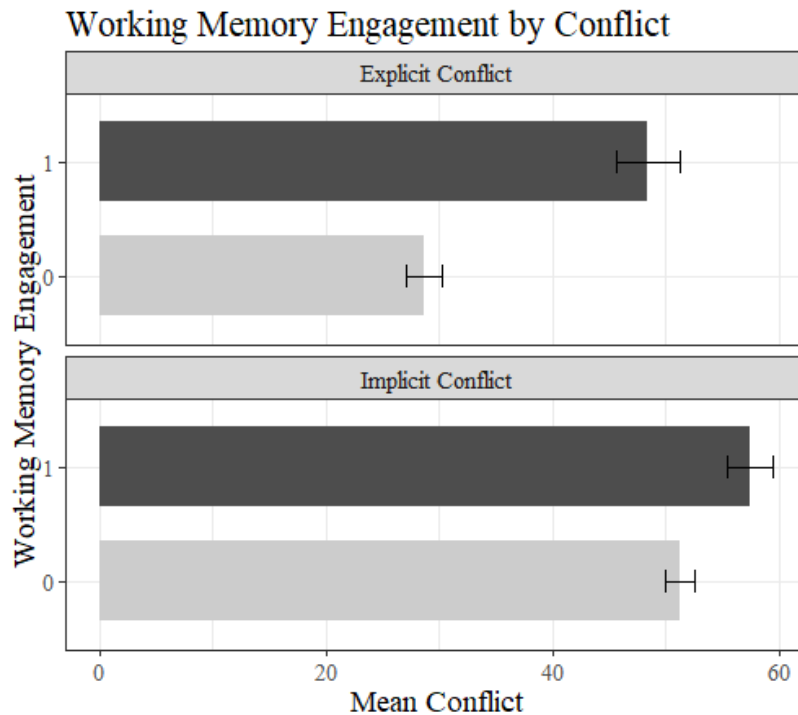


Figure 4. Working memory engagement by explicit and implicit conflict factors.

1=Working memory engaged, 0=Working memory not engaged. Error bars reflect +/- 1SE.

The relationship between explicit conflict, implicit conflict and working memory engagement

To examine whether explicit and implicit conflict were related we first examined their bivariate correlation which revealed that explicit and implicit conflict were not significantly associated, $r(767)=.063$, $p=.081$. However, earlier analyses suggested that both conflict factors were associated with working memory engagement. Therefore, we tested whether the two conflict factors made unique contributions to working memory engagement. The following analysis included both explicit and implicit measures of

conflict in the one binary logistic generalised linear mixed model predicting working memory engagement. The data was structured such that item was nested within test point, and test point within participant. The model revealed significant effects of item, $F(2, 363)=9.16$, $p<.001$, $\eta_p^2=.059$, and test point, $F(3, 363)=4.61$, $p=.001$, $\eta_p^2=.037$, on working memory engagement. As above, working memory engagement was more likely for items 1 and 3 than item 2 (see Appendix Table G), and decreasingly likely as training increased (see Appendix Table H). Interestingly, the model also demonstrated significant effects of both explicit conflict, $F(1,363)=20.28$, $\eta_p^2=.053$, $p<.001$, $OR=1.022$ [$CI=1.012,1.031$], and implicit conflict $F(1,363)=4.56$, $\eta_p^2=.012$, $p=.033$, $OR=4.394$ [$CI=1.124,17.167$] on working memory engagement. The results suggest that for both conflict factors, as conflict increased, the likelihood of working memory engagement increased. Furthermore, they suggest that, while controlling for the effects of test point and item, both conflict factors made a unique contribution to the model predicting working memory engagement.

Discussion

This study examined the proposition that uncertainty is involved in the engagement of reflective, Type 2 processes. By combining a cognitive constraint paradigm and measures of both explicit and implicit conflict, we were able to address three research questions. Firstly, we examined whether explicit and implicit conflict occurred during responding on the CRT and whether conflict changed with increased training. We found evidence for explicit but not implicit conflict on HL compared to CNL trials. There was no evidence that explicit or implicit conflict changed with increased training. Secondly, we examined whether explicit and implicit conflict predicted working memory engagement and whether this effect changed across test point. Our findings suggested that both explicit and implicit conflict predicted working memory engagement;

however, there was no evidence that this relationship changed across test point. Finally, we found that explicit and implicit conflict measures were not correlated, and that they each made unique contributions to the engagement of working memory.

Previous studies of conflict have found evidence to support the notion that the CRT activates conflicting logical principles that affect Type 1 processing. These experiments typically employ bias tasks (i.e. those that commonly elicit incorrect heuristic responses) to examine whether heuristic reasoners are sensitive to the task's conflicting logical principles. For example, De Neys et al. (2013) and Hoover and Healy (2019) found participants reported lower confidence on lure items than no-lure items; and Bago et al. (2019) found better-than-chance responding on second guess attempts at the bat and ball problem. However, Hoover and Healy (2019) and Purcell, Howarth, et al. (n.d.⁶⁰) conjectured that different measures of conflict may reveal different psychological phenomena. Therefore, the current study used two measures of conflict, one confidence-based and another gaze-based.

In line with previous studies, participants in the current study reported lower confidence on HL trials than CNL trials (De Neys et al., 2013; Purcell, Howarth, et al., n.d.⁶¹). However, in contrast to Purcell, Howarth, et al. (n.d.⁶²) there was no evidence that implicit (gaze-based) conflict differed between HL and CNL trials. These results suggest that the explicit measure is capturing a difference between these trial types, but the implicit measure was not. This is an intriguing finding. Taken at face value, it suggests that our confidence ratings are impacted by the conflicting mathematical principles, but our pattern of fixations is not. It is possible that the participants had some awareness of their error, but that this was driven by an explicit phenomenon, for example, the

^{60, 61, 62} Paper 3 in the current thesis.

participants may have learned which questions they were likely to get wrong but did not possess sufficiently intuitive logical principles to elicit implicit conflict. Alternatively, it could be an artefact of the procedure. The fixation measures were taken at Response 1 which had a time-limit imposed. The implicit conflict score was derived from the frequency of fixations on each multiple-choice alternative which could have been reduced or skewed by the limit time that the participants had to review the options. Thereby, it could be that the use of a time-limit at Response 1 may have artificially reduced the participants' implicit conflict scores. That implicit conflict was predictive of working memory engagement in later analyses suggested that while this measure was capturing some working memory related construct, future studies should clarify the utility of fixation-based implicit conflict measures under timed conditions.

Cognitive constraint studies have found evidence to support the suggestion that conflict is associated with working memory engagement. For example, using base-rate and syllogistic tasks, Bago and De Neys (2017) found that confidence ratings were lower and reaction times were longer for trials in which participants changed their answer from Response 1 to Response 2. However, they did not examine whether those measures made unique contributions to the engagement of working memory (i.e. answer change). Consistent with previous findings, the current study suggested that both explicit and implicit conflict were related to response change. Additionally, however, it suggested that explicit and implicit conflict factors made independent contributions to the engagement of working memory. This finding has implications for empirical and theoretical developments in thinking and reasoning research.

Previously, explicit indicators like confidence and feelings of error have been interpreted as indirect measures of conflict sensitivity occurring in Type 1 processing (e.g., De Neys et al., 2013; Gangemi et al., 2015). The interpretation of explicit conflict

measures as indicators of implicit conflict is challenged by the current finding that (confidence-based) and implicit (gaze-based) conflict were independent predictors of working memory engagement. One of the core functions of working memory is to facilitate the combination of different pieces of information (e.g., Baddeley, 1986). Therefore, it could be that indicators of explicit conflict are influenced by additional information or processes leading to deviations between implicit and explicit conflict indicators. The nature of that deviation presents an interesting avenue for future investigations into conflict and reasoning. Although the current findings must be explored further, particularly for replication and generalisation, they introduce the idea that models of reasoning may need to consider a continuum or multitude of conflict factors (i.e., ranging from implicit to explicit) and their potentially distinct roles in reasoning.

The current findings support the hybrid dual process models; however, their implications are slightly different for each model. For example, the metacognitive dual process model asserts that the fluency with which a response comes to mind impacts the reasoner's affect and in turn may cue working memory engagement. The current findings support this suggestion, in particular, it could be argued that implicit conflict may be an indicator of fluency, and explicit conflict an indicator of the subsequent affective response. Alternatively, Pennycook et al.'s (2015) three-stage dual process model suggests that multiple Type 1 processes can be triggered by a stimulus (stage-1), following which, conflict may be detected (stage-2). If the conflict is successfully detected, Type 2 processing is engaged (stage-3). Under this model, our finding that explicit conflict was registered for HL compared to CNL trials suggests that explicit conflict indicators may reflect Pennycook et al.'s stimulus-level stage-1 processes, whereas the explicit and implicit conflict indicators may reflect stage-2 processes (i.e., successful conflict detection and subsequent engagement of working memory). A more

comprehensive explanation of the current findings is afforded by the logical intuitions dual process model (De Neys, 2012).

The logical intuition dual process model asserts that multiple Type 1 processes can be engaged simultaneously, and that the process with the greatest activation is typically enacted (De Neys, 2012). The relative activation or salience of triggered Type 1 processes determines if cognitive conflict occurs; if one or more Type 1 processes have a similar level of activation, conflict is generated, which can lead to the engagement of Type 2, working memory dependent processes (De Neys, 2012, 2014). The current findings suggest that conflict predicts working memory engagement. This relation may reflect conflict occurring due to the similar activation of competing Type 1 processes and, subsequently, the engagement of Type 2 processes. The current study also demonstrated that as training increased, working memory engagement decreased. It could be that, with practice, the salience of the logical process (i.e. that leading to the correct solution) gradually increased. As the salience of the logical process increased, it may have surpassed the salience of the competing processing (i.e. that leading to the heuristic response) such that the relative salience of the two processes became increasingly unequal (see Purcell, Wastell, & Sweller, n.d.⁶³). In line with the current findings, greater practice should lead to increasingly unequal salience and decreasing Type 2 engagement. The current study supports the logical intuition model's assertion that Type 2 processes are be triggered by conflict and that conflict may be related to the reasoners' underlying process salience.

Broadly, the current findings support the hybrid dual process models, however, they also emphasise the need for clarification within those models of the proposed role of

⁶³ Paper 1 in the current thesis.

conflict as a continuous or multi-faceted phenomenon. The current investigation built on suggestions that the indicators of conflict may reflect distinct psychological processes (Hoover & Healy, 2019; Purcell, Howarth, et al., n.d.⁶⁴). It incorporated a training paradigm with the two-response method and two measures of conflict detection: confidence-based and gaze-based. As predicted by the hybrid dual process models, we demonstrated that the two measures were associated with working memory engagement. Additionally, we found that the explicit and implicit indicators made unique contributions to the model predicting working memory engagement. These findings deepen our understanding of *how* working memory is engaged and inform the continued development of theories of reasoning.

⁶⁴ Paper 3 in the current thesis.

References

- Baddeley, A. (1986). *Working memory*. Clarendon Press.
- 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*, 1–43.
<https://doi.org/10.1080/13546783.2018.1507949>
- Bago, B., Raelison, 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>
- Browne, M., Pennycook, G., Goodwin, B., & McHenry, M. (2014). Reflective minds and open hearts: Cognitive style and personality predict religiosity and spiritual thinking in a community sample. *European Journal of Social Psychology*, *44*(7), 736–742.
<https://doi.org/10.1002/ejsp.2059>
- De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. Retrieved from <http://pps.sagepub.com>
- 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., 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., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*(2), 269–

273. <https://doi.org/10.3758/s13423-013-0384-5>
- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102(2), 211–245. <https://doi.org/10.1037/0033-295X.102.2.211>
- 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>
- Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology (2006)*, 71(5), 1188–1208. <https://doi.org/10.1080/17470218.2017.1313283>
- 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>
- Hambrick, D. Z., & Engle, R. W. (2003). The role of working memory in problem solving. *The Psychology of Problem Solving*, 176–206.
- Hambrick, D. Z., & Oswald, F. L. (2005). Does domain knowledge moderate involvement of working memory capacity in higher-level cognition? A test of three models. *Journal of Memory and Language*, 52(3), 377–397. <https://doi.org/10.1016/j.jml.2005.01.004>
- Hoover, J. D., & Healy, A. F. (2019). The bat-and-ball problem: Stronger evidence in support of a conscious error process. *Decision*. Washington, DC : <https://doi.org/10.1037/dec0000107>
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: 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: Cambridge University Press.
- 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>
- Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review*, 19(3), 528–534. <https://doi.org/10.3758/s13423-012-0249-3>
- Purcell, Z. A., Howarth, S., Wastell, C. A., & Sweller, N. (n.d.). Strategy and conflict on the Cognitive Reflections Test: An eye tracking study. [Paper 1 in the current thesis]
- Purcell, Z. A., Wastell, C. A., & Sweller, N. (n.d.). Domain-Specific Experience and Dual-Process Thinking. [Paper 2 in the current thesis]
- Purcell, Z. A., Wastell, C. A., Sweller, N., & Howarth, S. (n.d.). No Pain, No Gain: Does Cognitive Conflict Predict Learning and Effortful Reasoning? [Paper 3 in the current thesis]
- Thompson, V. A. (2013). Why It Matters: The Implications of Autonomous Processes for Dual Process Theories-Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, 8(3), 253–256. <https://doi.org/10.1177/1745691613483476>
- Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, 20(2), 215–244. <https://doi.org/10.1080/13546783.2013.869763>
- 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>
- Thompson, V., & 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>
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289. <https://doi.org/10.3758/s13421-011-0104-1>
- 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>

Appendix

Table A. *Examples of training items and feedback.*

Training Item	Incorrect Feedback
A pen and a notebook cost \$25 in total. The notebook costs \$5 more than the pen. How much does the pen cost?	<p>* Sorry, your answer was incorrect. Don't worry, we'll help you find the correct solution!</p> <p>The question stated: <i>A pen and a notebook cost \$25 in total. The notebook costs \$5 more than the pen. How much does the pen cost?</i></p> <p>It's often helpful to breakdown these problems into an algebraic format. Try and give it go!</p> <p>The question states "A pen and a notebook cost \$25 in total."</p> <p>How can we write this algebraically?</p> <p>Hint: let p represent "pen" and n represent "notebook"</p> <p>g) $p + n = 25$ h) $p - n = 25$ i) $p + 25 = n$</p>
If it takes 6 adults 6 minutes to blow up 6 balloons, how long would it take 12 adults to blow up 12 balloons?	<p>Sorry! Your response was incorrect.</p> <p>The original question stated: <i>If it takes 6 adults 6 minutes to blow up 6 balloons, how long would it take 12 adults to blow up 12 balloons?</i></p> <p>The correct answer is 6 minutes. Each person can blow up ONE balloon in 6 minutes, you have 12 adults now and over 6 minutes each of them blows up ONE balloon. There are 12 adults, so they can blow up 12 balloons in total.</p> <p>Try the next question.</p>
In the desert, there is an anthill. Every day, the anthill doubles in height. If it takes 4 days for the anthill to reach 2 meters tall, long would it take for the anthill to reach 1 meter tall?	<p>Sorry! Your response was incorrect.</p> <p>The question stated: <i>In the desert, there is an anthill. Every day, the anthill doubles in height. If it takes 4 days for the anthill to reach 2 meters tall, long would it take for the anthill to reach 1 meter tall?</i></p> <p>The correct answer is 3 days. If on the third day the anthill is 1 meter tall, and it doubles in height every day, the anthill will be 2 meters tall on the fourth day.</p> <p>See if you can get the next one!</p>

Notes. *Feedback for training items related to CRT item 1 include several

algebraic steps. For brevity, they are not included but are available from the authors on request.

Table B. *Linear mixed model for the analysis on implicit conflict with effects of item, test point, AOI and problem type on conflict.*

Source	Df _{Source}	Df _{Error}	F	η_p^2	p
Item	2	974.05	.05	<.001	.950
Test Point	3	938.29	.04	<.001	.991
Trial Type	1	1108.91	.02	<.001	.892
AOI	3	1592.58	58.59	.099	<.001
Test Point * Trial Type	3	1099.23	.05	<.001	.985
Test Point * AOI	9	1608.24	.77	.004	.648

Trial Type * AOI	3	1602.23	52.69	.090	<.001
Test Point * Trial Type * AOI	9	1605.02	.52	.003	.864

Note: *Df Error* = 3294

Table C. *Follow up analysis for the effect of item within the binary logistic generalised linear mixed model predicting working memory engagement.*

Comparison	Coefficient	SE	<i>t</i>	OR	CI	<i>p</i>
1 - 2	-1.699	.38	-4.43	.183	.086, .389	<.001
1 - 3	-.559	.31	-1.83	.572	.314, 1.042	.068
2 - 3	1.124	.40	2.81	3.078	1.400, 6.768	.005

Note. *Df error* = 364. 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).

Table D. *Follow up analysis for the effect of test point within the binary logistic generalised linear mixed model predicting working memory engagement.*

Comparison	Coefficient	SE	<i>t</i>	OR	CI	<i>p</i>
1 - 2	-.349	.35	-1.00	.705	.356, 1.398	.316
1 - 3	-.864	.36	-2.44	.422	.208, .854	.017
1 - 4	-1.328	.39	-3.42	.265	.123, .569	.001
2 - 3	-.523	.38	-1.37	.593	.279, 1.257	.172
2 - 4	-.965	.41	-2.349	.381	.170, .854	.019
3 - 4	-.438	.42	-1.03	.645	.280, 1.487	.303

Note. *Df error* = 364. Comparisons are coded according to the test points (e.g., 1 = test point 1).

Table E. *Follow up analysis for the effect of item within the binary logistic generalised linear mixed model predicting working memory engagement.*

Comparison	Coefficient	SE	<i>t</i>	OR	CI	<i>p</i>
------------	-------------	----	----------	----	----	----------

1 - 2	-1.811	.38	-4.76	.163	.077, .345	<.001
1 - 3	-.515	.30	-1.72	.597	.332, 1.076	.086
2 - 3	1.294	.40	3.26	3.646	1.672, 7.951	.001

Note. *Df error* = 364. 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).

Table F. Follow up analysis for the effect of test point within the binary logistic generalised linear mixed model predicting working memory engagement.

Comparison	Coefficient	SE	<i>t</i>	OR	CI	<i>p</i>
1 - 2	-.371	.34	-1.10	.690	.356, 1.337	.271
1 - 3	-.840	.34	-2.46	.432	.221, .845	.014
1 - 4	-1.511	.38	-4.02	.221	.105, .462	<.001
2 - 3	-.480	.36	-1.32	.619	.303, 1.264	.187
2 - 4	-1.129	.40	-2.85	.323	.149, .705	.005
3 - 4	-.644	.41	-1.59	.525	.237, 1.164	.112

Note. *Df error* = 364. Comparisons are coded according to the test points (e.g., 1 = test point 1).

Table G. Follow up analysis for the effect of item within the binary logistic generalised linear mixed model predicting working memory engagement.

Comparison	Coefficient	SE	<i>t</i>	OR	CI	<i>p</i>
1 - 2	-1.650	.39	-4.250	.192	.090, .412	<.001
1 - 3	-.552	.31	-1.773	.576	.312, 1.062	.077
2 - 3	1.080	.4037	2.674	2.944	1.331, 6.513	.008

Note. *Df error* = 364. 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).

Table H. *Follow up analysis for the effect of test point within the binary logistic generalised linear mixed model predicting working memory engagement.*

Comparison	Coefficient	SE	<i>t</i>	OR	CI	<i>p</i>
1 - 2	-.356	.35	-1.01	.700	.350, 1.403	.314
1 - 3	-.852	.36	-2.36	.426	.209, .869	.019
1 - 4	-1.348	.39	-3.44	.260	.120, .561	.001
2 - 3	-.500	.3861	-1.295	.606	.284, 1.296	.196
2 - 4	-.976	.4145	-2.353	.377	.167, .852	.019
3 - 4	-.472	.4263	-1.107	.624	.270, 1.443	.269

Note. *Df error = 364. Comparisons are coded according to the test points (e.g., 1 = test point 1).*

General Discussion

Research Aims and Primary Findings

This thesis examined two core aspects of dual process models of reasoning (e.g., Evans & Stanovich, 2013; Thompson, Turner & Pennycook, 2011; De Neys, 2012, 2014). First, it aimed to test the assertion that, through increases in domain-specific experience, reasoning processes can become automatic such that they can operate without working memory engagement. Second, it aimed to test the hypothesis that cognitive conflict, generated by the relative activations of intuitions is associated with working memory engagement (see also Bago & De Neys, 2017). Across six experimental studies, this thesis found partial support for these hypotheses.

Study 1 of Paper 1 employed a between-subjects examination of the automisation hypothesis. Participants were recruited from populations that reflect low, intermediate and high differences in real-world mathematical experience. They were randomly allocated to the cognitive load or no load conditions and completed the original form of the CRT (Frederick, 2005). As expected, the low experience group performed poorly regardless of load; the intermediate group's performance was greater with no load than load; and the high experience group performed well regardless of load. This differential effect of load on performance depending on experience supports the automisation hypothesis, which states that working memory dependence should decrease once the reasoner is able to reach the correct solution and that solution process is practiced. However, there were a number of potential confounds in the study, for example, the decreased effect of the working memory load for high experience compared to intermediate experience participants may reflect differences in working memory capacity rather than differences in domain-specific experience. To address this issue, a second study was conducted.

Study 2 of Paper 1 used a within-subjects training manipulation to examine the automisation hypothesis. Participants were undergraduate psychology students – the same population as the low mathematical experience group in Study 1. They were randomly

allocated to one of three conditions: low, medium, and high load. They completed the original three-item form of the CRT at three test points. Feedback and guidance on CRT-like items was provided between the test points. Performance on the original CRT items improved as training increased. The effect of load condition differed across test point and working memory capacity. The effect of load was greater for participants with low working memory capacity than those with high working memory capacity. Of those with low working memory capacity, the effect of load was greater at the intermediate point of testing, than at the initial or final test point. This supported the results in the first study, suggesting that as training increased, the dependency of participants on working memory to complete the CRT problems decreased. However, the original three CRT items were employed at each test point and therefore, the reduction in working memory dependence could reflect rote learning rather than domain-specific experience based automatization effects. Although this is a possibility, participants were not provided with the answers to the original CRT items so they would have had to recall the answers from their previous working, which seems unlikely. Nonetheless, the remaining four studies presented in this thesis employed CRT-like problems rather than the original items.

Paper 2 included two studies that also used within-subject training manipulations, in this case, with new versions of the questions at each test point that were generated to reflect the structure of the three CRT items. Study 1 employed a matrix memory task as in Paper 1. In contrast to Paper 1, there was no evidence of the expected interactive effects of load constraint, domain-specific experience and working memory capacity on performance. Several methodological differences could explain the inconsistency between the results in Paper 1 (Study 2) and Paper 2 (Study 1). One possible reason for the inconsistency stems from the administration of the task. The study in Paper 1 (Study 2) was conducted face-to-face, whereas, the study in Paper 2 (Study 1) was conducted online. Therefore, we could not be sure that the participants in Paper 2 (Study 1) completed the

secondary task in line with the instructions, for example they could have used memory aids. Relatedly, these participants were recruited from the general public and so were likely to vary on a multitude of individual difference factors like age and intelligence. These differences may have influenced the individual's trajectories of automatisisation, thereby altering their patterns of working memory dependence. To address the first concern, a second study was conducted which employed a timing constraint rather than a visual memory task constraint.

In Paper 2, Study 2, a two-response paradigm was employed such that participants completed each CRT-like problem twice; initially with five seconds to respond, and again with no time-limit. In contrast to our hypotheses, there was no evidence for an interaction between response (constrained or unconstrained), domain-specific experience, and working memory capacity on performance as observed in Paper 1 (Study 2). Unlike Study 1 of Paper 2, this was not likely to result from participants not following instructions because the timing constraint could not be circumvented. Therefore, it seems more likely that the differences between Study 2 of Paper 1, and the studies in Paper 2 that lead to the inconsistent findings may stem from the breadth of individual differences contributing to noise in the data.

Another methodological difference that may account for the inconsistency between Paper 1 (Study 2) and Paper 2 (Studies 1 and 2) is the versions of the CRT that were employed. In Paper 1 the original three CRT items were administered, whereas in the studies in Paper 1 alternate versions of the CRT problems were employed. The original CRT items contain simple numbers like \$1.10 and \$1; generating alternate versions necessarily resulted in the use of more complex numbers. While the training paradigm may have been sufficient to elicit automatisisation with the use of problems requiring simple arithmetic, as observed in Paper 1, it may not have been sufficient for automatisisation on problems requiring more difficult calculations. Although the paradigm did not demonstrate

evidence for the automisation hypothesis, the training manipulation lead to a lower stability index. That is, most participants completed trials on which they demonstrated working memory engagement, and trials on which they did not demonstrate working memory engagement. This facilitated a more rigorous examination of the second core assumption in this thesis: that conflict is associated with working memory engagement.

In addition to testing the automisation hypothesis, Study 2 of Paper 2 also examined the relationship between conflict and working memory engagement. Participants completed both lure and no lure versions of the CRT-like problems and reported their confidence in their answers. Recall that lure items contain cues for two solutions, a heuristic and a logical response, whereas no lure items cue only a single solution. Conflict scores were calculated by subtracting participants' confidence ratings on the lure items from their confidence ratings on the corresponding no lure items. As in previous studies, answer change between participants' initial and final responses was interpreted as an indication of working memory engagement. The results of the study indicated that participants were more likely to engage working memory on trials for which they reported greater conflict. This lends support to the assertion that conflict is associated with working memory engagement. However, conflict is asserted to be the result of implicit, Type 1 processes having similar levels of activation. Therefore, the remaining two studies included an additional, implicit measure of conflict.

Paper 3 examined two aspects of reasoning on the CRT; first, the strategies participants used when completing the CRT, and second, the conflict experienced by heuristic responders. The study in Paper 3 was similar to those in Papers 1 and 2, however, the CRT-like problems were presented in a four-option multiple-choice format which allowed us to observe the participants' consideration of the alternatives. Travers, Rolison, and Feeney (2016) used a similar format with mouse-tracking and found that participants who gave the correct response spent more time considering the heuristic response than

participants who gave the heuristic response spent considering the correct response. They asserted that this finding supported the dual-process interpretation of strategy on the CRT; that correct responders may initially consider the heuristic response and then override this response. In contrast, Paper 3 did not observe differences between correct- and heuristic-respondents in the proportion of consideration of the heuristic and correct responses, respectively. In other words, in Paper 3, there was no evidence that participants were employing default-intervention strategies.

Paper 3 also included an examination of participants' implicit and explicit conflict experienced while solving the CRT. Implicit conflict was assessed using two new techniques based on the number of fixations and duration of dwell recorded for each multiple-choice option. The greater the spread of fixations and dwell between the multiple-choice options, the greater the implicit conflict. As in previous conflict studies, heuristic responding on the lure items was compared to correct responding on the no lure items. The analysis revealed that implicit conflict was greater on lure than no lure items for both fixation-based, and dwell-based implicit measures. Additionally, there was evidence for explicit, confidence-based conflict. That is, explicit conflict was greater on lure than no lure items. The implicit measures of conflict were highly correlated with each other, as anticipated, but intriguingly they were only weakly correlated with the explicit measure of conflict. Therefore, in the final Paper we examined whether these factors make independent contributions to the engagement of working memory.

Paper 4 reflects a culmination of the primary findings in the preceding papers. The study in Paper 4 employed a two-response paradigm in conjunction with a training manipulation of domain-specific experience, and included both implicit and explicit measures of conflict. In using a two-response paradigm we were able to create a variable for working memory engagement (as in Paper 2, Study 2) based on whether the participant had changed their response from their initial attempt to their second, final, attempt. In

employing a training manipulation, we elicited a low stability index, thereby increasing the credibility of the within-subject examination of conflict and its effects. Including both implicit and explicit measures of conflict facilitated the analysis of their potentially separate contributions to the engagement of working memory. The findings from the study in Paper 4 suggested that implicit and explicit conflict made unique contributions to the model predicting working memory engagement. This finding stands outside of the thesis' original predictions, which were largely based on the assumptions of the logical intuition model. The implications of this finding, and those in the preceding papers, for the logical intuition model are discussed below.

Implications for the logical intuition model

This thesis examined two core assertions in the logical intuition model. The first assertion, known as the “automisation assumption”, postulates that Type 2 processes can become Type 1 processes with practice. That is, as domain-specific experience increases, working memory dependence decreases. This was examined to some extent in all six studies but most directly by the studies in Paper 1. The second assertion addressed in this thesis postulates that conflict, determined by the relative strengths of intuitions, is associated with working memory engagement. This was examined to some extent in Papers 2, 3 and 4, but most directly in Paper 4.

The findings in Paper 1 supported the automisation hypothesis. They demonstrated differential dependence on working memory for the successful completion of the CRT problems as domain-specific experience increased. The findings support the suggestion of a parabolic relationship between domain-specific experience, working memory engagement, and performance. That is, low domain-specific experience was associated with heuristic responding and low working memory engagement, in contrast, intermediate domain-specific experience was associated with correct but working memory dependent responding, and high domain-specific experience was associated with correct responding

with low working memory dependence. This can be interpreted from a dual process perspective as a transition from incorrect Type 1 processing, to correct Type 2 processing, and finally, correct Type 1 responding. Broadly, it supports the suggestion that Type 2 processes can become Type 1 processes with practice.

Under the logical intuition dual process model, the findings from Paper 1 can be interpreted as an indication that the strength of the process corresponding to the normatively logical response increased with practice. The logical intuition model asserts that the process with the strongest activation will typically “win out”. That is, the process with the strongest activation is reflected by the response given by the participant. The improvement in performance observed in all six studies in this thesis lends credence to the idea that DSE increased the activation of the logical response. However, this not a controversial finding as most models of reasoning can account for improved performance with training. The findings in regard to the automisation hypothesis that assist in differentiating between models of reasoning are those suggesting a parabolic pattern of working memory engagement over increases in domain-specific experience.

While the logical intuition model asserts that the process with the greatest activation will typically be carried out, there is an important caveat. The model suggests that if there is another process with a similar level of activation, working memory may be engaged and an alternative solution may be enacted. From the logical intuition model’s perspective, the fact that in all studies in the thesis participants' performance improved after practice suggests that the absolute strength of the logical intuition eventually surpassed that of the heuristic intuition. For this to occur, assuming that the improved performance reflected a linear increase in the strength of the intuition, there must be some point along the trajectory of increasing domain-specific experience at which the heuristic and logical processes have a similar level of activation. At that point of surpassing, or near surpassing, the logical intuition model would predict higher levels of working memory

engagement.

In Paper 1, we observed greater working memory engagement at intermediate points of domain-specific experience than at low and high levels. Interpreted from the logical intuition model perspective, this finding suggests that the point of surpassing may have been captured by the studies' test points in the studies in Paper 1 corresponding to the "intermediate levels" of the domain-specific experience manipulations. This lends significant support to the logical intuition model and its shared dual process assumption that Type 2 processes can become Type 1 processes with practice. More specifically, this observation supports the assertion that the relative strengths of activations determine the engagement of working memory. However, it does rest on the assumption that the strength of a process is experience-dependent. While that is the approach in the current thesis, there may be additional or alternative factors that contribute to the strength of a process such as motivation or familiarity. Investigating other factors that may contribute to the strength of a process could be an interesting avenue for future research.

The second assertion of the logical intuition model that is examined in this thesis is that conflict is associated with working memory engagement. This was examined in Papers 2, 3, and 4. In Study 1 of Paper 2, we found that explicit conflict became a stronger predictor of performance as domain-specific experience increased. In Study 2 of Paper 2, using a two-response paradigm, we demonstrated that explicit conflict was associated with working memory engagement. However, as highlighted in the General Introduction and in Papers 3 and 4, conflict is thought to stem from the competition between intuitive (Type 1) processes. Moreover, the confidence-based conflict factors might reflect other phenomena not related to the underlying strengths of activations. For example, particularly given that these paradigms included training with feedback and guidance, it could be that the participants were learning which types of problems they were likely to get right or wrong, rather than learning the correct procedures and changing the underlying processes or

activations. Therefore, we aimed to examine whether the effect of conflict was also observed using more implicit measures.

Papers 3 and 4 explored the possibility that conflict could be measured using more implicit eye-tracking tools. In Paper 3, evidence for implicit conflict was demonstrated using measures of fixations and dwell across the multiple-choice options. That is, the fixation- and dwell-based measures showed significant differences between heuristic-lure responding and correct-no lure responding. These measures were correlated with an explicit, confidence-based measure of conflict which supported the validity of the measures. However, the correlation between implicit and explicit measures was weak. This was the first result in the thesis to suggest that these tools may be measuring different phenomena.

In Paper 4, a two-response paradigm was employed. In contrast to Paper 3, the comparison of heuristic responding on lure items and correct responding on no lure items did not reveal significant differences in the implicit, fixation-based measure. Recall that during the recording of eye movements, participants in Paper 3 were reasoning under cognitive load constraints (low or high) with no timing restrictions while Paper 4 imposed a time-based restraint at the initial response. Therefore, it may be that the implicit conflict factor was a more effective measure of conflict in the paradigm with no time-limit imposed. That is, the implicit conflict measure was effective without a timing-constraint (Paper 3) but not with time-constraint (Paper 4). The implicit measure was derived from the frequency with which the participants looked to each multiple-choice response option. By limiting the amount of time respondents could take to consider the possible options, we may have artificially reduced the participants' conflict scores. In contrast to the implicit conflict factors, the explicit conflict factor showed significant differences between heuristic-lure responses and correct-no lure response in both paradigms. This was the second indication throughout the thesis that the implicit and explicit measures may reflect

different aspects of conflict.

In Paper 4, once again, answer change from initial to second response was interpreted as an indication of working memory engagement. When examined independently, both implicit and explicit conflict factors were predictive of working memory engagement. However, they were not significantly correlated with one another. Therefore, both factors were entered into the model predicting working memory engagement. Both factors accounted for a significant portion of variance in working memory engagement. This was the third indication that implicit and explicit conflict may reflect different phenomena.

There are several explanations that may account for the independent contributions of implicit and explicit conflict to working memory engagement. First, the implicit factor may reflect the underlying differences in reasoning processes – potentially the competition between Type 1 processes, whereas the explicit factor may be a symptom of that conflict, exacerbated, for example, by the increased cognitive resources allocated to mental representations that gain access to working memory and the global workspace. Second, the two types of conflict may reflect different influences on the reasoning process, for example, while the implicit factor may tap into the conflict generated by intuitions with similar levels of activation, the explicit factor may reflect the anticipation of positive or negative feedback. Certainly, there are more possible explanations of the differences between implicit and explicit indicators of conflict. The possibility of separate sources of conflict requires greater empirical examination, pending which, the findings in regard to implicit and explicit conflict may present a new element for consideration within the logical intuition model's current hypothesis for conflict and working memory engagement.

Alternative Explanations

The findings from this thesis are interpreted primarily in relation to the logical intuition dual process model, however, there are other prominent models that should be

considered. The metacognitive reasoning theory posits that, given consistent cognitive capacity and environmental constraints, the extent to which a person engages in Type 2 thinking is determined by metacognitive factors (Thompson et al., 2011; Thompson, 2009). The model suggests that a third set of processes monitor the Type 1 processes, acting as a trigger for Type 2 processes to be engaged. Monitoring is described, as it is in the metamemory literature, as a “subjective assessment of one’s own cognitive processes and knowledge” (Koriat, Ma’ayan & Nussinson, 2006, p. 38). Type 1 processes are thought to produce two outputs: The content of the initial response and an accompanying metacognitive assessment of the correctness of that response (Thompson et al., 2011; Thompson, 2009). The metacognitive assessment is determined by various implicit cues such as the ease with which a response is brought to mind and is referred to as a “feeling of rightness” (Thompson et al., 2011).

The metacognitive model can be used to explain the relationship between “conflict”, as the term is employed in the current thesis, and Type 2 engagement. The findings reported in Papers 4 and 5, in particular, can be readily explained by the metacognitive model. The studies in Papers 4 and 5 include measures of conflict obtained through eye-tracking and self-reported confidence ratings. The eye-tracking data suggests that uncertainty, determined by the proportion of fixations upon the non-selected response options, predicts working memory engagement. This is in line with the metacognitive and logical intuition models of reasoning. However, the patterns of fixations were not in line with the logical intuition model’s suggestion that conflict results from the competition between two Type 1 processes.

Under the logical intuition model, we would expect the participants to look at the response options that reflect the two competing Type 1 processes. In the case of the bat and ball problem, for example, these responses would be 5c (corresponding to the logical response) and 10c (corresponding to the heuristic response). In Papers 4 and 5, the

questions were presented to the participants with four response options from which they could select, for example, 5c, 10c, 15c, 1c. The results indicate that the more the participants looked at the alternatives, the more likely they were to change their response, indicating further reasoning and working memory engagement. However, the responses that the participants looked at were not in line with the expectation from the logical intuition model. That is, there were more fixations on the non-selected responses, but these fixations did not occur on the options corresponding to the potentially competing processes. Rather, they were evenly spread between the three non-selected responses. This indicates that uncertainty was occurring, and predictive of working memory engagement, but that this uncertainty may not stem from the competition between two competing processes. Therefore, it may be argued that, particularly for the findings in Papers 4 and 5, the metacognitive model may offer an explanation that is as good as, if not better than that offered by the logical intuition model. Although the aim of the current thesis was to examine the automatisisation and conflict from the perspective of the logical intuition model, the findings could be applied to other hybrid dual process models, in particular, the metacognitive model. Building on the current findings, future studies could generate *a priori* hypotheses to directly compare the claims about “feeling of rightness” and “conflict” as postulated in the metacognitive and logical intuition models.

Strengths, weaknesses, and future directions

This project highlights several avenues for future research for to validate and extend the current findings. Although the individual papers delve into specific recommendations to extend and validate the studies within each article, it is important to consider the strengths and weaknesses of the thesis as a whole. Therefore, the following section examines the overarching methodological aspects of the project – the use of different cognitive constraints, power and effect sizes, and the CRT. It then explores opportunities for extending the current findings through the development of new

approaches and areas of investigation by integrating the current findings with individual difference research or initiating a deeper investigation into conflict.

Methodological issues.

In six studies, the CRT (or CRT-like problems) were administered in conjunction with one of two methods for cognitive constraint. Cognitive constraints were employed to experimentally reduce the likelihood of Type 2 processing by decreasing the participants' temporary cognitive capacity. This was achieved by imposing a matrix memory task (Papers 1, 2, and 3) or a timing constraint within a two-response paradigm (Paper 4). These techniques yielded slightly different effects across the studies. In Paper 1, the matrix memory load was effective in demonstrating differences in working memory engagement across domain-specific experience. However, in Paper 1, Study 2, the pattern of decreased performance was such that the 'high' load was less detrimental than the 'medium' load. There could be several reasons for this, for example the 'high' load patterns may have formed more recognisable patterns that made for easier recall. For the remaining matrix memory constraint studies, matrix patterns were employed that reflected the 'low' and 'medium' load patterns in Paper 1, Study 2. However, the use of the matrix memory task was not effective in the remaining studies. In Paper 2, it was thought that the online nature of the study may have reduced the effect of the load. However, in Paper 3's study, conducted face-to-face, the matrix task was once again ineffective.

The matrix task was effective in the first two studies in which it was used (Paper 1, Studies 1 and 2) but not the latter two studies (Paper 2, Study 1, and Paper 3). The grid patterns in the earlier administration were presented for 800 ms, whereas, they were presented for 900 ms in the later administration. Matrix tasks have been effectively administered for different amounts of time in the past, for example, De Neys and Verschueren (2006) presented the a grid for 850 ms whereas (Johnson, Tubau, & De Neys, (2016) presented similar grids for 900 ms. However, future studies may benefit from

considering these factors more closely before administering matrix memory tasks as a form of cognitive constraint, particularly when used in conjunction with the CRT.

The current thesis employed two-response paradigms in Papers 2 and 4. As such it was able to investigate the role of conflict and working memory engagement as operationalised through timing constraints. However, recent studies have employed both load and timing constraints in two-response paradigms to increase the likelihood that the initial responses were likely to reflect Type 1 processing (e.g., Bago & De Neys, 2017, 2019). To validate the current findings, future studies could use a similar approach, simultaneously employing memory load and timing constraints in a two-response paradigm in conjunction with manipulations of domain-specific experience.

The presence of both significant and null findings for similar hypotheses may stem from methodological differences, as in the case of the matrix presentation, or from situational differences, as in the case of online versus offline administrations. However, they may also stem from issues surrounding power and effect sizes. There were two primary lines of empirical enquiry warranting further evaluation in regard to power and effect sizes. The first focused on the three-way interaction between constraint, WMC and DSE on performance (examined in three studies: Paper 1, Study 2; Paper 2, Study 1; and Paper 3) and the second focused on the effect of conflict on working memory engagement (examined in four studies: Paper 2, Study 1 and 2, and Paper 3 and 4).

The first enquiry yielded significant results in Paper 1 (Study 2, $\eta^2_p = .093$, $N=80$) and non-significant results in Papers 2 (Study 1, $\eta^2_p = .009$, $N=107$) and 3 ($\eta^2_p = .072$, $N=38$). The effect size in Paper 2 was smaller than those in Papers 1 and 3. However, the sample size was larger than that in Paper 1 and, therefore, sufficient to detect the effect demonstrated in Paper 1. Therefore, the non-significance in Paper 2 is more likely to reflect methodological rather than power issues. Specifically, Paper 2 studies were administered online using MTurk participants whereas Papers 1 and 3 studies were

administered offline using psychology undergraduate students. The effect sizes in Paper 1 and Paper 3 were comparable, however, the sample size in Paper 3 was less than half that in Paper 1. The non-significant result in Paper 3, therefore, may be a result of low statistical power. Considered together, these studies suggest that the three-way interaction between DSE, WMC, and constraint will replicate in offline studies with sufficient power to detect effect sizes of approximately $\eta^2_p = .070$. It is important to note that the sample sizes in Papers 3 and 4 were determined by the primary aims of the papers—examining the relationship between conflict and working memory—not the relationship between constraint, WMC and DSE.

The second enquiry, regarding conflict and working memory engagement, was examined in Papers 2 (Study 1 and 2), 3 and 4. All four studies examined explicit conflict (via confidence ratings) and Papers 3 and 4, additionally, examined implicit conflict (via eye tracking). In Paper 2, Study 1 (N=100), we hypothesised that constraint, conflict and DSE would affect performance. However, constraint—a manipulation to determine working memory engagement via a visual memory task—did not contribute to any significant effects in the model: main effect ($\eta^2_p = .001$), two-way (constraint*DSE, $\eta^2_p = .001$; constraint*conflict, $\eta^2_p = .001$) or three-way interaction ($\eta^2_p = .005$). These small effect sizes suggest that these non-significant results stem from an ineffective constraint manipulation rather than a lack of power. The lack of constraint effects yielded Paper 2, Study 1 ultimately unhelpful for examining the relationship between conflict and WME. However, Paper 2 Study 2 (N=107) demonstrated an effect of constraint using an alternative, time-limit based working memory engagement manipulation.

Paper 2, Study 2 found that constraint had a significant, small-medium effect on performance ($\eta^2_p = .086$) but no evidence that this effect differed across DSE (two-way interaction between constraint and DSE: $\eta^2_p = .002$). A significant effect of constraint meant that the data from Study 2 was appropriate for examining the relationship between conflict

and working memory engagement. To account for the time-varying conflict factor, binary dependent variable and nested data structure, a binary logistic generalised linear model was employed to test if conflict predicted working memory engagement. The analysis revealed a significant medium-large effect of conflict on working memory engagement ($\eta^2_p = .123$). This indicates that the sample size, $N=121$, was sufficient for detecting the effects of the primary variables of interest: constraint (working memory engagement) and conflict.

Papers 3 and 4 examined explicit and implicit conflict, measured via confidence ratings and eye movements respectively. The sample sizes in Papers 3 ($N=38$) and 4 ($N=38$) ensured that they were powered to find an effect size of $\eta^2_p = .123$, based on the preceding study of explicit conflict (Paper 2, Study 2). Paper 3 found evidence for explicit conflict on heuristic-lure (HL) compared to correct-no lure (CNL) items. This was indicated most directly by the significant effect of problem type (HL, CNL) on explicit conflict ($\eta^2_p = .165$; see Paper 3, Appendix C, Table E). The study also demonstrated evidence of implicit conflict. This was indicated most directly by the effect of the interaction between AOI and problem type (HL, CNL) on fixation count ($\eta^2_p = .007$; see Paper 3, Appendix C, Table F). Paper 4 used these indicators of conflict to examine the relationship between conflict and working memory engagement. Both explicit ($\eta^2_p = .053$) and implicit conflict ($\eta^2_p = .013$) were significant predictors of working memory engagement with small to medium effect sizes. This indicates that the study was sufficiently powered for its primary research analyses. It should be noted that the power for this study and many other in the thesis was assisted by the within-subject manipulations of DSE via training and working memory engagement via time-limit based constraint methods. Although the effect sizes indicate meaningful effects across both lines of enquiry, the need to replicate those effects is paramount. In particular, future studies should aim to examine whether the three way interaction observed in Paper 1 Study 2 can be replicated.

This project involves six studies that employ the three item CRT or versions of the

three item CRT that retain the structure of the problems but not the content or quantities.

The CRT has been used extensively in reasoning and bias research, however, many studies include only the bat and ball problem (or versions of this problem; e.g., Hoover & Healy, 2019; Szollosi, Bago, Szaszi, & Aczel, 2017). Using the three-item version of the test allows for the examination of reasoning on all three items, and also increases the validity of the assessment. Interestingly, when including item in the analyses for Paper 4, working memory engagement was more likely to occur for items 1 and 3 than item 2. This suggests that automatisisation may be faster for item 2 than the other items, or that item 2 requires less working memory in general. The possibility that the items do not exhibit the same pattern of working memory engagement may have implications for the administration of the test in other contexts, for example, as a measure of reflection or cognitive ability. Future studies should be conducted to directly examine the potentially different dependence on working memory for the three items.

Using the three-item version of the CRT increases the validity of the findings to an extent, however, employing a test with only three items is still a small number of items from which to draw conclusions about the participants' reasoning processes and about reasoning more generally. It will be important for future studies to examine whether the current findings generalise to other bias tasks such as base-rate neglect problems and syllogistic reasoning tasks (see General Introduction for examples). Moreover, it would be interesting to examine the trajectory of working memory engagement in domains outside of mathematics. However, the examination of changes in mathematical experience and its effect on reasoning processes, as employed in the current project, contains several outstanding questions. In particular, the individual differences that may affect the automatisisation of mathematical reasoning.

Integrating individual differences.

The six studies in the current thesis use the differences in working memory

engagement elicited by changes in domain-specific experience to examine the interaction between Type 1 and Type 2 processing. The studies in Paper 1 focussed on the trajectory of automatisisation, however, the subsequent studies used the shifts in working memory dependence to ensure lower stability indices and generate a more rigorous examination of conflict. The hypothesised parabolic relationship between domain-specific experience and working memory engagement was supported by the findings in Paper 1. Although not their primary aim, the subsequent articles did not find a consistent pattern of working memory engagement across changes in domain-specific experience. This could be due to individual differences in, for example, cognitive abilities such as intelligence, numeracy, or working memory capacity, or cognitive styles such as open-mindedness or need for cognition (Stanovich, 2009). It is likely that differences in these factors would lead to differences in a person's trajectory of automatisisation. For example, a person with higher intelligence or need for cognition may exhibit a faster process of automatisisation. Moreover, a participant with high numeracy may demonstrate signs of automatisisation earlier than one with lower numeracy. Although controlling numeracy and working memory capacity was sufficient to demonstrate the expected relationship between experience and working memory engagement in Paper 1 (Study 2), future studies are needed to investigate the role of these factors in automatisisation in more detail and incorporate additional individual difference measures.

The future of conflict.

The results pertaining to the differences between implicit and explicit indicators of conflict suggest that one of the most important aims for future research will be to gain a deeper understanding of conflict, its measurement, underlying mechanisms and consequences for reasoning. The examination of conflict within reasoning research is a rapidly growing area and, accordingly, there are an increasing number of techniques used to measure conflict. Conflict has been measured using techniques such as autonomic

arousal (De Neys, Moyens, & Ansteenwegen, 2010), brain imaging (Simon, Lubin, Houdé, & De Neys, 2015), ‘feelings of rightness’ and confidence (De Neys, Rossi, & Houdé, 2013). Although the relationship between different measurements, like confidence and response times, have been used to support the validity of those techniques (e.g., Frey, Johnson, & De Neys, 2017; Hoover & Healy, 2019), the consideration that they may reflect separate phenomena or make unique contributions to the engagement of working memory engagement has not been directly examined.

Although the current project builds on previous theorising that has used the implicit and explicit distinction to differentiate between measurement types (De Neys, Vartanian, & Goel, 2008), this dichotomy should be considered with caution. The current findings suggest that the implicit and explicit measurements of conflict may make unique contributions to the engagement of working memory. However, future studies are needed to examine whether they are measuring unique aspects of the same underlying construct, or whether they are measuring different constructs that independently contribute to working memory engagement. To address this, future studies would need to determine if double dissociative effects can be obtained using these factors. Currently, it cannot not be concluded that the tools used to measure conflict in the current project reflect separate phenomena. Therefore, care should be taken when interpreting the current results. The current findings do not, in and of themselves, support a dualistic model of conflict (implicit and explicit), however, they do present sufficient evidence to prompt further investigation.

The continued investigation into the role of metacognitive factors such as conflict in reasoning should be conducted with consideration of the relationship between the terms used to label various kinds of cognitive uncertainty. The terms ‘conflict’, ‘feeling of rightness’, ‘confidence’ and ‘uncertainty’ are often used interchangeably and this thesis has, largely, followed that trend. However, while cognitive conflict can be interpreted generally to convey that a reasoner is subjectively conflicted about their response, it can

also be interpreted directly from the logical intuition model that refers to the conflict between two or more Type 1 processes. It should be noted that the findings in the current thesis do not necessarily support the latter, direct, interpretation and further research is needed to test the mechanism underlying indications of conflict as proposed by the logical intuition model. The indirect measurements of conflict, for example via confidence ratings, were predictive of Type 2 processing. This is in line with the logical intuition model but it depends on the assumption that confidence is an indirect measure of conflict. Confidence ratings may be affected by other factors such as fluency as the metacognitive model would suggest. The continued exploration of metacognitive factors and their underlying mechanism may help in distinguishing between the contemporary hybrid models. However, to achieve that, the field requires significant semantic clarifications of the terms mentioned above and, importantly, a systematic clarification of the ways in which they can be operationalised.

Conclusion

This thesis contributes to our understanding of the processes that underlie thinking and reasoning by examining how reasoning that is initially effortful can become intuitive and automatic. This phenomenon is universal and – perhaps deceptively – simple; indeed, it is the very purpose of many everyday pursuits. However, while readily acknowledged by many scholars, it has received surprisingly little empirical consideration within the field of reasoning. The aim of the current thesis was, first, to examine the assumption that a reasoner can transition from using Type 2, effortful reasoning processes that require working memory to solve a problem, to using automatic Type 1 reasoning processes for the same type of problem, and second, to harness this transition to examine the relationship between Type 1 and Type 2 thinking. While the first aim focused on this relationship from the perspective of the decreasing engagement of working memory associated with learning, the second aim examined the same relationship to determine when individuals improved

their performance on a reasoning problem by changing from using automatic Type 1 processing to using working memory-dependent Type 2 processing. The project examined how a reasoner might eventually automate the problem-solving process to the extent that it no longer requires working memory resources and can be successfully solved using intuitive Type 1 processes, and when and why a reasoner might engage Type 2 processes to solve a problem.

This was achieved in six studies within four papers that examined the nature of people's reasoning on the CRT. In Paper 1 the "automatisation hypothesis" was supported in a between-subjects comparison of participants with different levels of real-world mathematical experience, and again in a within-subjects examination of participants' reasoning throughout a newly developed training paradigm. Participants with greater domain specific experience were able to use Type 1 processing to solve CRT problems. In Papers 2 to 4, the training paradigm was adapted and combined with measures of cognitive conflict to examine the logical-intuition model of reasoning's assertion that conflict is associated with the engagement of working memory. Evidence was found in support of this hypothesis and results also highlighted the possibility that conflict may not be a unitary construct.

The thesis adopted the idea that the "strength" or salience of a particular reasoning process can be conceptualised as determined by an individual's domain-specific experience. By employing this definition, clear a priori hypotheses could be formed around the assertions made in the logical intuition dual process model about conflict and working memory engagement. These were formed into four research questions by separately examining the relationships between performance on the CRT, working memory engagement, conflict and domain-specific experience (see General Introduction, Table 3). Examining the relationship between domain-specific experience and performance on the CRT, we found support for the suggestion that a process' absolute salience increases with

experience. Testing the relationship between conflict and performance, we found evidence to suggest that cognitive conflict changes with experience such that it becomes a stronger predictor of performance. Examining the relationship between domain-specific experience and working memory engagement, we found evidence that successful completion of the CRT is less demanding on working memory as mathematical experience increases. However, while lowered stability indices (induced via the training paradigm) allowed for the testing of the thesis' hypotheses, the pattern of that engagement was not consistent across studies, which suggests that other factors, possibly individual differences, may affect individuals' trajectories of automatisisation. Finally, testing the relationship between conflict and working memory engagement, we observed a positive relationship between both implicit and explicit measures of conflict and working memory engagement.

Taken together, these results support the assertion that increasing domain specific experience can increase the strength of a particular (correct) reasoning process such that it conflicts with other (incorrect) reasoning processes. This increased relative strength of the correct process causes cognitive conflict, but also causes the reasoner to engage working memory and reason more effortfully to successfully solve the problem. When the correct reasoning process has become sufficiently stronger than the competing incorrect processes, less cognitive conflict is generated, which signals that working memory is no longer required to solve this type of problem, at which point the correct response can be considered to be automatised. This thesis supports the core assumptions in the logical intuition model and presents an empirical argument for a detailed integration of domain-specific experience into the model.

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*, 1–43.
<https://doi.org/10.1080/13546783.2018.1507949>
- De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. Retrieved from <http://pps.sagepub.com>
- 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., Moyens, E., & Ansteenwegen, D. V. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective and Behavioral Neuroscience*, 10(2), 208–216. <https://doi.org/10.3758/CABN.10.2.208>
- De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, 20(2), 269–273. <https://doi.org/10.3758/s13423-013-0384-5>
- De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science*. New York, NY :
<https://doi.org/10.1111/j.1467-9280.2008.02113.x>
- De Neys, W., & Verschueren, N. (2006). Working Memory Capacity and a Notorious Brain Teaser. *Experimental Psychology*. Göttingen, Germany :
<https://doi.org/10.1027/1618-3169.53.1.123>
- Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>

Frey, D., Johnson, E. D., & De Neys, W. (2017). Individual differences in conflict detection during reasoning. *The Quarterly Journal of Experimental Psychology*, 1–52.

<https://doi.org/10.1080/17470218.2017.1313283>

Hoover, J. D., & Healy, A. F. (2019). The bat-and-ball problem: Stronger evidence in support of a conscious error process. *Decision*. Washington, DC :

<https://doi.org/10.1037/dec0000107>

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, 164, 56–64.

<https://doi.org/10.1016/j.actpsy.2015.12.008>

Simon, G., Lubin, A., Houdé, O., & De Neys, W. (2015). Anterior cingulate cortex and intuitive bias detection during number conservation. *Cognitive Neuroscience*. Hove, East Sussex : <https://doi.org/10.1080/17588928.2015.1036847>

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

Szollosi, A., Bago, B., Szaszi, B., & Aczel, B. (2017). Exploring the determinants of confidence in the bat-and-ball problem. *Acta Psychologica*, 180, 1–7.

<https://doi.org/10.1016/J.ACTPSY.2017.08.003>

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>

Ethics Approval

Ethics Approval of this thesis has been removed as it may contain sensitive/confidential content

Appendix

Questions developed for Paper 1 (Study 2)***Example questions:***

1. A flower and a vase cost \$10 in total. The vase costs \$7. How much does the flower cost? ____ dollars
2. It takes John 4 minutes to run 6 miles. How long would it take for him to run 12 miles? ____ minutes
3. Sally borrowed 5 books. She returned 3. How many does she have left? ____ books

Test questions (the original CRT; Frederick, 2005):

1. A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost? ____ cents (A: 5 cents)
2. It takes 5 machines 5 minutes to make 5 widgets. How long would it take 100 machines to make 100 widgets? ____ minutes (A: 5 minutes)
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. It takes 48 days for the patch to cover the entire lake. How long would it take for the patch to cover half of the lake? ____ days (A: 47 days)

Training questions:

1. A purse and a bag cost \$9 in total. The bag costs \$1 more than the purse. How much does the purse cost? ____ dollars

Feedback (Correct):

Well done! Your answer was correct.

Did you know you can use algebra to solve these types of questions?

For example, if you had a question:

“A lamp and a lamp shade cost \$100 in total. The lamp costs \$40 more than the shade.

How much does the shade cost?”

You could rewrite it as:

$$(1) \quad l + s = 100$$

$$(2) \quad l - s = 40$$

$$(3) \quad s = ?$$

You can then use substitution to solve for s . First, rearrange (2) for l : $l = s + 40$

Second, substitute this equation into (1): $(s + 40) + s = 100$

Then solve for s : $2s + 40 = 100$, $2s = 100 - 40$, $2s = 60$, $s = 30$

Here, we've use algebra to solve problem and work out that the lamp shade costs \$30.

If you are finding this hard to follow, use your notepad to write out the steps. These will be useful for the next question.

Keep up the good work!

Feedback (Incorrect):

Sorry, your answer was incorrect. Don't worry, we'll help you reach the correct solution!

The question stated: A purse and a bag cost \$9 in total. The bag costs \$1 more than the purse. How much does the purse cost?

It's often helpful to breakdown these problems into an algebraic format. Try and give it go!

The question states "A purse and a bag cost \$9 in total." How can we write this

algebraically? *Hint: let p represent "purse" and b represent "bag"*

(a) $p+9=b$

(b) $p-b=9$

(c) $p+b=9$

The question also states "The bag costs \$1 more than the purse." How can we write this

algebraically? *Hint: let p represent "purse" and b represent "bag"*

(a) $b-p=1$

(b) $b+1=p$

(c) $b+p=1$

[Next page]

The question stated: A purse and a bag cost \$9 in total. The bag costs \$1 more than the purse. How much does the purse cost?

From the question, we know that:

(1) $p + b = 9$

(2) $b - p = 1.$

Let's combine these statements to find p .

First, take statement (2) and rearrange for b . How can you rewrite (2) to put b on left-hand side of the equation?

(a) $b=9+1$

(b) $b=1+p$

(c) $b = 1 - p$

Next, let's substitute our new equation for b into statement (1). You can write this:

(a) $p + (1 + p) = 9$

(b) $(1 + p) - p = 1$

(c) $b = 1 + (b - 1)$

[Next page]

The question stated: A purse and a bag cost \$9 in total. The bag costs \$1 more than the purse. How much does the purse cost?

You've worked out that: $p + (1 + p) = 9$

Let's solve for p ! First, let's make that equation simpler. How can we rewrite this equation?

(a) $2p + 1 = 9$

$2p = 8$

(b) $2p + 1 = 9$

$2p = 10$

(c) $p + 1 = 9 + p$

$p = 8 + p$

Can you work out the value of the purse?

(a) The purse is \$5

(b) The purse is \$4

(c) The purse is \$8

[Next page]

Using a similar process, can you solve the question:

A purse and a bag cost \$13 in total. The bag costs \$1 more than the purse. How much does the purse cost?

Hint: Try following these steps: 1. Write "A purse and a bag cost \$13" algebraically. 2.

Write "The bag costs a dollar more than the purse" algebraically. 3. Rearrange (2) so the b is on the left hand side. 4. Substitute this equation into the statement from step 1. 5.

Simplify and solve for p !

(a) The purse costs \$12

(b) The purse costs \$6

(c) The purse costs \$4

[Next page (correct)]

Congratulations! Your answer is correct!

[Next page (incorrect)]

Your answer was not correct. Have another try!

A purse and a bag cost \$13 in total. The bag costs \$1 more than the purse. How much does the purse cost?

Hint: Try following these steps: 1. Write "A purse and a bag cost \$13" algebraically. 2.

Write "The bag costs a dollar more than the purse" algebraically. 3. Rearrange (2) so the b is on the left-hand side. 4. Substitute this equation into the statement from step 1. 5.

Simplify and solve for p !

(a) The purse costs \$6

(b) The purse costs \$12

(c) The purse costs \$4

[Next page]

2. If it takes 4 people 3 minutes to peel 4 potatoes, how long would it take 9 people to peel 9 potatoes? ____ minutes

Feedback (correct):

Congratulations! Your answer was correct!

The question stated: If it takes 4 people 3 minutes to peel 4 potatoes, how long would it take 9 people to peel 9 potatoes?

The correct answer is 3 minutes. Each person can peel ONE potato in 3 minutes, you have 9 people working, and so over 3 minutes each of them peels ONE potato. There are nine people so together they can peel 9 potatoes in total.

Feedback (incorrect):

Sorry! Your response was incorrect.

The question stated: If it takes 4 people 3 minutes to peel 4 potatoes, how long would it take 9 people to peel 9 potatoes?

The correct answer is 3 minutes. Each person can peel ONE potato in 3 minutes, you have 9 people working, and so over 3 minutes each of them peels ONE potato. There are nine people so together they can peel 9 potatoes in total.

[Next page]

3. In the desert, there is an anthill. Every day, the anthill doubles in height. If it takes 4 days for the anthill to reach 2 meters tall, long would it take for the anthill to reach 1 meter tall? ____ days

Feedback (incorrect):

Sorry! Your response was incorrect.

The question stated: In the desert, there is an anthill. Every day, the anthill doubles in height. If it takes 4 days for the anthill to reach 2 meters tall, long would it take for the anthill to reach 1 meter tall?

The correct answer is 3 days. If on the third day the anthill is 1-meter-tall, and it doubles in height every day, the anthill will be 2 meters tall on the fourth day.

See if you can get the next one!

Feedback (correct):

Congratulations! Your response was correct.

The question stated: In the desert, there is an anthill. Every day, the anthill doubles in height. If it takes 4 days for the anthill to reach 2 meters tall, long would it take for the anthill to reach 1 meter tall?

The correct answer is 3 days. If on the third day the anthill is 1-meter-tall, and it doubles in height every day, the anthill will be 2 meters tall on the fourth day.

See if you can get the next one!

[Next page]

4. A stick of butter and a bottle of milk cost \$1.20 in total. The butter costs \$1 more than the milk. How much does the milk cost? ____ cents

Feedback (correct):

Well done! Your answer was correct.

Did you know you can use algebra to solve these types of questions?

For example, if you had a question:

A lamp and a lamp shade cost \$100 in total. The lamp costs \$40 more than the shade. How much does the shade cost?

You could rewrite it as: (1) $l + s = 100$ (2) $l - s = 40$ (3) $s = ?$

You can then use substitution to solve for s . First, rearrange (2) for l : $l = s + 40$

Second, substitute this equation into (1): $(s + 40) + s = 100$

Then solve for s : $2s + 40 = 100$ $2s = 100 - 40$ $2s = 60$ $s = 30$

Here, we've use algebra to solve problem and work out that the lamp shape costs \$30.

If you are finding this hard to follow, use your notepad to write out the steps. These will be useful for the next question.

Keep up the good work!

Feedback (incorrect):

Sorry, your answer was incorrect. Don't worry, we'll help you reach the solution!

The original question stated: A stick of butter and a bottle of milk cost \$1.20 in total. The butter costs \$1 more than the milk. How much does the milk cost?

It's often helpful to breakdown these problems into an algebraic format. Try and give it go!

The question states "A stick of butter and a bottle of milk cost \$1.20 in total" How can

we write this algebraically? *Hint: let b represent "butter" and m represent "milk"*

a) $b + m = 1.20$

b) $b - m = 1.20$

c) $b = 1.20 + m$

The question also states "The butter costs a dollar more than the milk". How can we write this algebraically? *Hint: let b represent "butter" and m represent "milk"*

a) $b - m = 1.00$

b) $b + 1.00 = m$

c) $b + m = 1.00$

[Next page]

The original question stated: A stick of butter and a bottle of milk cost \$1.20 in total. The butter costs \$1 more than the milk. How much does the milk cost?

From the question, we know that:

(1) $b + m = 1.20$

(2) $b - m = 1.00$

Let's combine these statements to find m .

First, take statement (2) and rearrange for b . How can you rewrite (2) to put b on left-hand side of the equation?

a) $m = b + 1.00$

b) $b = m + 1.20$

c) $b = 1.00 + m$

Next, let's substitute our new equation for b into statement (1). You can write this:

a) $b = 1.00 + m$

b) $(1.00 + m) + m = 1.20$

c) $(1.00 + m) - m = 1.00$

[Next page]

The original question stated: A stick of butter and a bottle of milk cost \$1.20 in total. The butter costs \$1 more than the milk. How much does the milk cost?

You've worked out that: $(1.00 + m) + m = 1.20$

Let's solve for m ! First, let's make that equation simpler. How can we rewrite this equation?

a) $1.00 + m = 1.00$ $m = 1.00$

b) $1.00 + 2m = 1.20$ $2m = 0.20$

c) $1.00 + 2m = 1.00$ $2m = 2.00$

Can you work out the value of the bottle of milk?

a) The bottle of milk is \$1.00

b) The bottle of milk is \$0.10

c) The bottle of milk is \$0.20

[Next page]

Using a similar process, can you solve the question:

A stick of butter and a bottle of milk cost \$2.40 in total. The butter costs \$1 more than the milk. How much does the milk cost?

Hint: Try following these steps: 1. Write "A stick of butter and a bottle of milk cost \$2.40" algebraically. 2. Write "The butter costs a dollar more than the milk" algebraically. 3. Rearrange the equation from step 2 so the b is on the left-hand side. 4. Substitute this equation into the statement from step 1. 5. Simplify and solve for m !

- a) The bottle of milk is \$1.40
- b) The bottle of milk is \$0.70
- c) The bottle of milk is \$1.20

[Next page (correct)]

Congratulations! Your answer is correct!

[Next page (incorrect)]

Sorry! Your answer was incorrect, have another go!

A stick of butter and a bottle of milk cost \$2.40 in total. The butter costs \$1 more than the milk. How much does the milk cost?

Hint: Try following these steps: 1. Write "A stick of butter and a bottle of milk cost \$2.40" algebraically. 2. Write "The butter costs \$1 more than the milk" algebraically. 3. Rearrange the equation from step 2 so the b is on the left-hand side. 4. Substitute this equation into the statement from step 1. 5. Simplify and solve for m !

- a) The bottle of milk is \$1.20
- b) The bottle of milk is \$1.40
- c) The bottle of milk is \$0.70

5. If it takes 2 elves 6 minutes to make 2 toys, how long would it take 50 elves to make 50 toys? ____ minutes

Feedback (correct):

Congratulations your answer is correct!

The question stated: If it takes 2 elves 6 minutes to make 2 toys, how long would it take 50 elves to make 50 toys?

The correct answer is 6 minutes. Each elf can make ONE toy in 6 minutes. There are 50 elves. So over 6 minutes, they can make 50 toys!

Try the next question.

Feedback (incorrect):

Sorry, your answer was incorrect.

The question stated: If it takes 2 elves 6 minutes to make 2 toys, how long would it take 50 elves to make 50 toys?

The correct answer is 6 minutes. Each elf can make ONE toy in 6 minutes. There are 50 elves. So over 6 minutes, they can make 50 toys!

Try the next question.

6. On a slice of bread, there is a patch of mould. Every day, the patch of mould doubles in size. If it takes 8 days for the mould to cover the whole slice, how long would it take for the mould to cover half the slice? ____ days

Feedback (correct):

Congratulations! Your response was correct!

The question stated: On a slice of bread, there is a patch of mould. Every day, the patch of mould doubles in size. If it takes 8 days for the mould to cover the whole slice, how long would it take for the mould to cover half the slice?

The correct answer is 7 days. If on the seventh day the mould had covered half the slice, and it doubles in size every day, the whole slice would be covered on the eighth day.

See if you can get the next one!

Feedback (incorrect):

Sorry! Your response was incorrect.

The question stated: On a slice of bread, there is a patch of mould. Every day, the patch of mould doubles in size. If it takes 8 days for the mould to cover the whole slice, how long would it take for the mould to cover half the slice?

The correct answer is 7 days. If on the seventh day the mould had covered half the slice, and it doubles in size every day, the whole slice would be covered on the eighth day.

See if you can get the next one!

[Next page]

7. A laptop with a warranty costs \$1100 in total. The laptop costs \$1000 more than the warranty. How much does the warranty cost? ____ dollars

Feedback (correct):

Well done! Your answer was correct.

Did you know you can use algebra to solve these types of questions?

For example, if you had a question:

A lamp and a lamp shade cost \$100 in total. The lamp costs \$40 more than the shade. How much does the shade cost?

You could rewrite it as:

$$(1) \quad l + s = 100$$

$$(2) \quad l - s = 40$$

$$(3) \quad s = ?$$

You can then use substitution to solve for s.

First, rearrange (2) for l: $l = s + 40$

Second, substitute this equation into (1): $(s + 40) + s = 100$

Then solve for s:

$$2s + 40 = 100$$

$$2s = 100 - 40$$

$$2s = 60$$

$$s = 30$$

Here, we've use algebra to solve problem and work out that the lamp shade costs \$30.

If you are finding this hard to follow, use your notepad to write out the steps. These will be useful for the next question.

Keep up the good work!

Feedback (incorrect):

Sorry, your answer was incorrect. Don't worry, we'll help you reach the solution!

The question stated: A laptop with a warranty costs \$1100 in total. The laptop costs \$1000 more than the warranty. How much does the warranty cost?

It's often helpful to breakdown these problems into an algebraic format. Try and give it go!

The question states "A laptop with a warranty costs \$1100 in total" How can we write

this algebraically? *Hint: let l represent "laptop" and w represent "warranty"*

- a) $l - w = 1100$
- b) $l + w = 1100$
- c) $l + 1100 = w$

The question also states "The laptop costs \$1000 more than the warranty"

How can we write this algebraically?

Hint: let l represent "laptop" and w represent "warranty"

- a) $l + w = 1000$
- b) $w - l = 1000$
- c) $l - w = 1000$

[Next Page]

The question stated: A laptop with a warranty costs \$1100 in total. The laptop costs \$1000 more than the warranty. How much does the warranty cost?

From the question, we know that: (1) $l + w = 1100$ (2) $l - w = 1000$

Let's combine these statements to find w .

First, take statement (2) and rearrange for l . How can you rewrite (2) to put l on left-hand side of the equation?

a) $l - 1000 = w$

b) $w = 1000 + l$

c) $l = 1000 + w$

Next, let's substitute our new equation for l into statement (1). You can write this:

a) $(1000 + w) + w = 1100$

b) $(1000 + w) + w = 1000$

c) $(1000 + w) - w = 1100$

[Next page]

The original question stated: A laptop with a warranty costs \$1100 in total. The laptop costs \$1000 more than the warranty. How much does the warranty cost?

You've worked out that: $(1000 + w) + w = 1100$

Let's keep simplifying! Which equations come next?

a) $1000 + 2w = 1100$

$$2w = 1100 - 1000$$

b) $1000 + 2w = 1000$

$$2w = 1000 - 1000$$

c) $1100 + 2w = 1000$

$$2w = 1000 - 1100$$

What is the value of the warranty?

a) The warranty is \$1000

b) The warranty is \$50

c) The warranty is \$100

[Next page]

Using a similar process, can you solve the question:

A laptop with a warranty costs \$1300 in total. The laptop costs \$1000 more than the warranty. How much does the warranty cost?

Hint: Try following these steps:

- 1. Write "A laptop with a warranty costs \$1300" algebraically.*
- 2. Write "The laptop costs \$1000 more than the warranty" algebraically.*
- 3. Rearrange the equation from step 2 so the l is on the left hand side.*
- 4. Substitute this equation into the statement from step 1.*
- 5. Simplify and solve for w !*

a) \$50

b) \$150

c) \$300

[Next page (correct)]

Congratulations! Your answer is correct!

[Next page (incorrect)]

Sorry, your answer was incorrect.

Have another go!

Using a similar process, can you solve the question:

A laptop with a warranty costs \$1300 in total. The laptop costs \$1000 more than the warranty. How much does the warranty cost?

Hint: Try following these steps: 1. Write "A laptop with a warranty costs \$1300"

algebraically. 2. Write "The laptop costs \$1000 more than the warranty" algebraically.

3. Rearrange the equation from step 2 so the l is on the left hand side. 4. Substitute this equation into the statement from step 1. 5. Simplify and solve for w !

a) \$150

b) \$300

c) \$50

[Next page]

8. If it takes 3 painters 5 hours to paint 3 walls, how long would it take 8 painters to paint 8 walls? ____ hours

Feedback (correct):

Congratulations! Your answer was correct!

The question stated: If it takes 3 painters 5 hours to paint 3 walls, how long would it take 8 painters to paint 8 walls?

The correct answer is 5 hours. Each painter can paint ONE wall in 5 hours, there are 8 painters working, and so over 5 hours they each paint ONE wall, in total they will have painted 8 walls.

See if you can get the next one.

Feedback (incorrect):

Sorry! Your answer was wrong.

The question stated: If it takes 3 painters 5 hours to paint 3 walls, how long would it take 8 painters to paint 8 walls?

The correct answer is 5 hours. Each painter can paint ONE wall in 5 hours, there are 8 painters working, and so over 5 hours they each paint ONE wall, in total they will have painted 8 walls.

See if you can get the next one.

9. A particular tree loses its leaves in Autumn every year. However, in Spring the leaves usually triple in number every day. On Day 4 of spring, the tree has around 600 leaves. On which day would you predict the tree to have had 200 leaves? Day

Feedback (correct):

Congratulations! Your answer was correct!

The question stated: A particular tree loses its leaves in Autumn every year. However, in Spring the leaves usually triple in number every day. On day 4 of spring, the tree has around 600 leaves. On which day would you predict the tree to have had 200 leaves? Day

The correct answer is Day 3. If, on Day 3, the tree had 200 leaves, the following day (Day 4), you would expect this number to TRIPLE. So, on Day 4, you would expect around 600 leaves.

Feedback (incorrect):

Sorry! Your answer was incorrect.

The question stated: A particular tree loses its leaves in Autumn every year. However, in Spring the leaves usually triple in number every day. On day 4 of spring, the tree has around 600 leaves. On which day would you predict the tree to have had 200 leaves? Day

The correct answer is Day 3. If, on Day 3, the tree had 200 leaves, the following day, you would expect this number to TRIPLE. So, on Day 4, you would expect around 600 leaves.

[Next page]

Well done! You have completed the first of two training blocks!

The following section includes THREE test items. Like the test items you completed earlier, these include grid patterns for you to memorise, the test question, and then a blank grid for you to recall the grid pattern. Remember, it is very important that you remember the grid pattern.

Training Block 2

The feedback for the items in training block 2 followed the same pattern as that in training block 1. Therefore, we have listed the remaining questions without the feedback.

1. A pen and a notebook cost \$25 in total. The notebook costs \$5 more than the pen.
How much does the pen cost? _____ dollars
2. If it takes 6 adults 6 minutes to blow up 6 balloons, how long would it take 12 adults to blow up 12 balloons? _____ minutes (A: 6)
3. A fast-spreading disease is affecting a high-density city. Every day that there is no cure available, the number of people infected doubles. It is expected to take 12 days

- for the disease to infect half the population of the city, how long would it you expect it to take for the disease to affect a quarter of the city's population? ____ days (A: 11)
4. The bills for electricity and gas is \$310 in total. The electricity costs \$10 more than the gas. How much does the gas cost? ____dollars (A: 150)
5. If it takes 15 assembly lines 20 days to build 15 cars, how long would it take 200 assembly lines to build 200 cars? ____days (A: 20)
6. A town's population is currently 15,000. If the population doubles every 55 years, what will the population be 110 years from now? ____people (A: 60'000)
7. A pair of sunglasses and a glasses case cost \$150 in total. The sunglasses cost \$100 more than the case. How much does the case cost? ____dollars (A: 25)
8. It takes 10 machines 15 minutes to make 100 chocolate bars. How long would it take 20 machines to make 200 chocolate bars? ____minutes (A: 15)
9. A scientist places 10 bacteria in a petri dish. The bacteria are expected to double in number every TWO days. How many days would it take until there are 40 bacteria? ____days (A: 4)

Questions developed for Papers 3 and 4

The questions used in papers 3 and 4 were presented in a multiple-choice format. The questions are listed below with the multiple-choice options.

Example questions:

1. A flower and a vase cost \$10 in total. The vase costs \$7. How much does the flower cost?

Correct: \$3

Incorrect-1: \$17

Incorrect-2: \$4

Incorrect-3: \$2

2. It takes John 4 minutes to run 6 miles. How many minutes would it take for him to run 12 miles?

Correct: \$8

Incorrect-1: \$4

Incorrect-2: \$12

Incorrect-3: \$16

3. Sally borrowed 5 books. She returned 3. How many does she have left?

Correct: \$2

Incorrect-1: \$3

Incorrect-2: \$8

Incorrect-3: \$12

Test questions (lure):

1. A bag and a badge cost \$12.10 in total. The bag costs \$2.00 more than the badge.

How much does the badge cost?

Correct: \$5.05

Incorrect-1: \$10.10

Incorrect-2: \$11.10

Incorrect-3: \$7.05

2. It takes 3 spiders 3 minutes to make 3 webs. How long would it take 100 spiders to make 100 webs?

Correct: 3 min

Incorrect-1: 100 min

Incorrect-2: 5 min

Incorrect-3: 10 min

3. In a garden, there is a patch of weeds. Every day, the patch doubles in size. It takes 28 days for the patch to cover the entire garden. How long would it take for the patch to cover half of the garden?

Correct: 27 days

Incorrect-1: 14 days

Incorrect-2: 20 days

Incorrect-3: 10 days

Test questions (no lure):

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

Correct: 90c

Incorrect-1: 45c

Incorrect-2: \$1.90

Incorrect-3: 60c

5. It takes 1 factory 10 days to build 20 cars. How many days would it take 1 factory to build 40 cars?

Correct: 20 days

Incorrect-1: 40 days

Incorrect-2: 60 days

Incorrect-3: 10 days

6. A school has 35 students. The number of students is expected to increase by 10 every year. How many years will it take for the school to have 55 students?

Correct: 2 years

Incorrect-1: 20 years

Incorrect-2: 10 years

Incorrect-3: 5years

Training questions (lure):

1. A purse and a bag cost \$9 in total. The bag costs \$1 more than the purse.

How much does the purse cost?

Correct: \$4

Incorrect-1: \$8

Incorrect-2: \$5

Incorrect-3: \$10

Feedback (correct):

One way you can solve this problem is by using algebra. Here is a quick step-by-step guide:

Step 1. Write “A purse and a bag cost \$9 in total” algebraically

$$p + b = 9$$

Step 2. Write “The bag costs \$1 more than the purse” algebraically

$$b - p = 1$$

Step 3. Then, take the statement from step 2 and rearrange for b.

$$b - p = 1$$

Step 4. Substitute the new equation for b into the statement from step 1.

$$p + (p + 1) = 9$$

Step 5. Rewrite the equation in step 4 to make it simpler.

$$2p + 1 = 9$$

$$2p = 9 - 1$$

$$2p = 8$$

Step 6. Solve p !

$$p = 4$$

Therefore, the purse is worth \$4!

Feedback (incorrect):

The question stated: *A purse and a bag cost \$9 in total. The bag costs \$1 more than the purse. How much does the purse cost?*

It's often helpful to breakdown these problems into an algebraic format. Try and give it go!

The question states "a purse and a bag cost \$9 in total."

How can we write this algebraically? Choose one of the following answers.

Hint: Let p represent "purse" and b represent "bag"

a) $p + b = 9$

b) $p - b = 9$

c) $p + 9 = b$

The question also states "The bag costs \$1 more than the purse."

How can we write this algebraically? Choose one of the following answers.

a) $b - p = 1$

b) $b + 1 = p$

c) $b + p = 1$

[Next page]

The question stated. A purse and a bag cost \$9 in total. The bag costs \$1 more than the purse. How much does the purse cost?

From the question (and our previous algebra), we know that:

1) $p + b = 9$

$$2) \quad b - p = 1$$

Let's combine these statements to find p .

First, take statement (2) and rearrange for b . How can you rewrite (2) to put b on the left-hand side of the equation? Choose one of the following answers.

$$a) \quad b = 1 + p$$

$$b) \quad b = 9 + 1$$

$$c) \quad b = 1 - p$$

Next, let's substitute our new equation for b into statement (1). Choose one of the following answers. You can write this:

$$a) \quad p + (1 + p) = 9$$

$$b) \quad (1 + p) - p = 1$$

$$c) \quad b = 1 + (b - 1)$$

[Next page]

The question stated: A purse and a bag cost \$9 in total. The bag costs \$1 more than the purse. How much does the purse cost?

You've worked out that:

$$p + (p + 1) = 9$$

Let's solve for p !

First, let's make that equation simpler. How can we rewrite this equation? Choose one of the following answers.

$$a) \quad 2p + 1 = 9$$

$$2p = 8$$

$$b) \quad p + 1 = 9 + p$$

$$p = 8 + p$$

$$c) \quad 2p + 1 = 9$$

$$2p = 10$$

Can you work out what the purse is worth?

- a) \$4
- b) \$8
- c) \$5

[Next page]

2. In 3 minutes 3 people peel 3 potatoes. How long would it take 9 people to peel 9 potatoes?

Correct: 3 minutes

Incorrect-1: 9 minutes

Incorrect-2: 4 minutes

Incorrect-3: 5 minutes

Feedback:

The question stated: In 3 minutes 3 people peel 3 potatoes. How long would it take 9 people to peel 9 potatoes? The correct answer is 3 minutes. Each person can peel ONE potato in 3 minutes, you have 9 people working, and so over 3 minutes each of them peels ONE potato. There are nine people so together they can peel 9 potatoes in total.

3. In the desert, there is a very large anthill. Every day, the anthill doubles in height. It takes 4 days for the anthill to reach 2 meters tall. How many days would it take for the anthill to reach 1 meter tall?

Correct: 3 days

Incorrect-1: 2 days

Incorrect-2: 4 days

Incorrect-3: 1 day

Feedback:

The question stated: In the desert, there is a very large anthill. Every day, the anthill doubles in height. It takes 4 days for the anthill to reach 2 meters tall. How many days would it take for the anthill to reach 1 meter tall? correct answer is 3 days. If on the third day the anthill is 1-meter-tall, and it doubles in height every day, the anthill will be 2 meters tall on the fourth day.

Training items (no lure):

4. The bills for electricity and gas are \$310 in total. The electricity costs \$110. How much does the gas cost?

Correct: \$200

Incorrect-1: \$100

Incorrect-2: \$210

Incorrect-3: \$110

Feedback:

The question stated: The bills for electricity and gas are \$310 in total. The electricity costs \$110. How much does the gas cost? The correct answer is \$200. You can work this out by subtracting the cost of the electricity from the total cost (i.e. $310 - 110 = 200$).

5. A typist can type 60 words per minute. How long would it take the typist to type 180 words?

Correct: 3 minutes

Incorrect-1: 9 minutes

Incorrect-2: 30 minutes

Incorrect-3: 6 minutes

Feedback:

The question stated: A typist can type 60 words per minute. How long would it take the typist to type 180 words? The correct answer is 3 minutes. If it takes 1 typist 1 minute to type 60 words, then you can work out the total number of minutes by dividing the total number of words (180) by the number of words per minute (60), meaning it would take the typist 3 minutes (i.e. $180/60 = 3$).

6. There are 12 people infected by a virus. The number of people infected is expected to increase by 10 per week. How many weeks will it take for the virus to infect 42 people?

Correct: 3 weeks

Incorrect-1: 1 week

Incorrect-2: 30 weeks

Incorrect-3: 2 weeks

Feedback:

The question stated: There are 12 people infected by a virus. The number of people infected is expected to increase by 10 per week. How many weeks will it take for the virus to infect 42 people? The correct answer is 3 weeks. If 10 more people are infected each week, and the starting number is 12 people, it would take 3 weeks for 42 people to be infected (i.e. $42 - 12 = 30$. $30/10 = 3$).

Test questions (lure):

1. A pen and a notebook cost \$25 in total. The notebook costs \$5 more than the pen.
How much does the pen cost?

Correct: \$10

Incorrect-1: \$20

Incorrect-2: \$15

Incorrect-3: \$25

2. If it takes 4 painters 4 hours to paint 4 walls. How many hours would it take 8 painters to paint 8 walls?

Correct: 4 hours

Incorrect-1: 8 hours

Incorrect-2: 2 hours

Incorrect-3: 16 hours

3. A tree loses its leaves in autumn every year. However, in spring, the leaves triple in number every day. After 6 days of spring, the tree has around 600 leaves. After how many days would the tree have had 200 leaves?

Correct: 5 days

Incorrect-1: 2 days

Incorrect-2: 4 days

Incorrect-3: 1 day

Test questions (no lure):

4. A water bottle and a mug cost \$22 in total. The water bottle costs \$15. How much does the mug cost?

Correct: \$7

Incorrect-1: \$5

Incorrect-2: \$10

Incorrect-3: \$15

5. It takes 1 chicken 1 day to lay 1 egg. How many days would it take 1 chicken to lay 5 eggs?

Correct: 5 days

Incorrect-1: 1 day

Incorrect-2: 3 days

Incorrect-3: 10 days

6. A new company has 10 employees. The number of employees is expected to increase by 2 every year. How many years will it take for the company to have 16 employees?

Correct: 3 years

Incorrect-1: 2 years

Incorrect-2: 10 years

Incorrect-3: 6 years

Training questions (lure):

1. A bubble gum packet and a mint cost \$1.20 in total. The bubble gum costs \$1 more than the mint. How much does the mint cost?

Correct: 10c

Incorrect-1: 20c

Incorrect-2: 5c

Incorrect-3: 15c

Feedback (correct):

The question stated:

A bubble gum packet and a mint cost \$1.20 in total. The bubble gum costs \$1 more than the mint. How much does the mint cost?

One way you can solve this problem is by using algebra. Here's a quick step-by-step guide:

Step 1: Write "A bubble gum packet and a mint cost \$1.20 in total." algebraically

$$b + m = 1.20$$

Step 2: Write "The bubble gum costs \$1 more than the mint." Algebraically

$$b = m + 1$$

Step 3: Then, take the statement from Step 2 and rearrange for b

$$b = m + 1$$

Step 4: Substitute the new equation for b in to the statement in Step 1

$$m + (m + 1) = 1.20$$

Step 5: Rewrite the equation in Step 4 to make it simpler

$$2m + 1 = 1.20$$

$$2m = 1.20 - 1$$

$$2m = 0.20$$

Step 6: Solve for m!

$$M = 10$$

Therefore, the mint is worth 10 cents!

Feedback (incorrect):

The question stated: A bubble gum packet and a mint cost \$1.20 in total. The bubble gum costs \$1 more than the mint. How much does the mint cost?

It's often helpful to break these questions down into an algebraic format. Try and give it a go!

The question states "A bubble gum packet and a mint cost \$1.20 in total."

How can we write this algebraically? Choose one of the following answers.

Hint: let b represent "bubble gum" and m represent "mint"

a) $m + b = 1.20$

b) $m - b = 1.20$

c) $m + 1.20 = b$

The question also states "the bubble gum costs \$1 more than the mint."

How can we write this algebraically? Choose from one of the following answers.

a) $b - m = 1$

b) $b + 1 = m$

c) $b + m = 1$

[Next page]

The question stated: A bubble gum packet and a mint cost \$1.20 in total. The bubble gum costs \$1 more than the mint. How much does the mint cost?

From the question (and our previous algebra), we know that:

1) $b + m = 1.20$

2) $b - m = 1$

Let's combine these statements to find b .

First, take statement (2) and rearrange for b . How can you rewrite (2) to put b on the left hand side? Choose from the following:

a) $b = 1 + m$

b) $b = 1 + 1.20$

c) $b = 1 - m$

Next, let's substitute our new equation for b into statement (1). Choose from the following:

a) $m + (m + 1) = 1.20$

b) $(1 + m) - m = 1$

c) $b = 1 + (b - 1)$

[Next page]

The question stated: A bubble gum packet and a mint cost \$1.20 in total. The bubble gum costs \$1 more than the mint. How much does the mint cost?

You've worked out that:

$$m + (m + 1) = 1.20$$

Let's solve for b!

First, let's make that equation simpler. How can we rewrite this equation? Choose from the following:

a) $2m + 1 = 1.20$

$$2m = 0.20$$

b) $m + 1 = 1.20 + b$

$$m + 1.20 + m$$

c) $2m - 1 = 1.20$

$$2m = -0.20$$

2. If it takes 25 elves 25 minutes to make 25 toys. How long would it take 50 elves to make 50 toys?

Correct: 25 minutes

Incorrect-1: 50 minutes

Incorrect-2: 20 minutes

Incorrect-3: 15 minutes

Feedback:

The question stated: If it takes 25 elves 25 minutes to make 25 toys. How long would it take 50 elves to make 50 toys? The correct answer is 25 minutes. Each elf can make one toy in 25 minutes, you have 50 elves working, and so over 25 minutes each of them makes ONE toy. There are 50 elves, so together they can make 50 toys in total.

3. A slice of bread has a patch of mould. Every day, the patch of mould triples in size. It takes 8 days for the mould to cover the whole slice. How many days would it take for the mould to cover a third of the slice?

Correct: 7 days

Incorrect-1: 4 days

Incorrect-2: 3 days

Incorrect-3: 2 days

Feedback:

The question stated: A slice of bread has a patch of mould. Every day, the patch of mould triples in size. It takes 8 days for the mould to cover the whole slice. How many days would it take for the mould to cover a third of the slice? The correct answer is 7 days. If on the seventh day, the mould covers a third of the slice of bread, and it triples in size every day, the mould will cover the whole slice of bread on the eighth day.

Training questions (no lure):

4. A pair of sunglasses and a glasses case cost \$150 in total. The sunglasses cost \$125. How much does the case cost?

Correct: \$25

Incorrect-1: \$125

Incorrect-2: \$50

Incorrect-3: \$100

Feedback:

The question stated: A pair of sunglasses and a glasses case cost \$150 in total. The sunglasses cost \$125. How much does the case cost? The correct answer is \$25. You can

work this out by subtracting the cost of the sunglasses from the total cost (i.e. \$150 - \$125 = \$25).

5. It takes 1 machine 15 minutes to make 100 chocolate bars. How long would it take 1 machine to make 200 chocolate bars?

Correct: 30 minutes

Incorrect-1: 15 minutes

Incorrect-2: 10 minutes

Incorrect-3: 100 minutes

Feedback:

The question stated: It takes 1 machine 15 minutes to make 100 chocolate bars. How long would it take 1 machine to make 200 chocolate bars? The correct answer is 30 minutes. If it takes 1 machine 15 minutes to make 100 chocolate bars, then it takes twice the amount of time to make 200 chocolate bars (i.e. $200/100 = 2$. $2 \times 15 = 30$).

6. A scientist places 10 bacteria in a petri dish. Each day this number increases by 5. How many days would it take until there are 20 bacteria?

Correct: 2 days

Incorrect-1: 10 days

Incorrect-2: 5 days

Incorrect-3: 4 days

Feedback:

The question stated: A scientist places 10 bacteria in a petri dish. Each day this number increases by 5. How many days would it take until there are 20 bacteria? The correct answer is 2 days. If the bacteria increased by 5 each day, and the bacteria started at 10, it would take 2 days for the bacteria to reach 20 (i.e. $20 - 10 = 10$. $10/5 = 2$).

Test questions (lure):

1. A bus ticket and a soda cost \$5.60 in total. The bus ticket costs \$1 more than the soda. How much does the soda cost?

Correct: \$2.30

Incorrect-1: \$4.60

Incorrect-2: \$3.60

Incorrect-3: \$3.30

2. It takes 10 children 10 minutes to read 10 books. How many minutes would it take 100 children to read 100 books?

Correct: 10 minutes

Incorrect-1: 100 minutes

Incorrect-2: 30 minutes

Incorrect-3: 50 minutes

3. In a field, there is a group of rabbits. Every month, the number of rabbits in the group doubles. After 6 months, the group has a total of 100 rabbits. How many months did it take for the number of rabbits to reach 50?

Correct: 5 months

Incorrect-1: 3 months

Incorrect-2: 4 months

Incorrect-3: 2 months

Test questions (no lure):

4. A massage and a manicure cost \$80 in total. The massage costs \$50. How much does the manicure cost?

Correct: \$30

Incorrect-1: \$20

Incorrect-2: \$80

Incorrect-3: \$40

5. It takes 1 runner 10 minutes to run 2 miles. How many minutes would it take 1 runner to run 4 miles?

Correct: 20 minutes

Incorrect-1: 10 minutes

Incorrect-2: 15 minutes

Incorrect-3: 30 minutes

6. A small zoo has 40 animals. The number of animals is expected to increase by 5 animals every month. How many months will it take for the zoo to have 45 animals?

Correct: 1 month

Incorrect-1: 5 months

Incorrect-2: 3 months

Incorrect-3: 2 months

Training questions (lure):

1. A coffee and a muffin cost \$7.40 in total. The muffin costs \$1 more than the coffee. How much does the coffee cost?

Correct: \$3.20

Incorrect-1: \$6.40

Incorrect-2: \$5.60

Incorrect-3: \$8.40

Feedback (correct):

The question stated: A coffee and a muffin cost \$7.40 in total. The muffin costs \$1 more than the coffee. How much does the coffee cost?

One way you can solve this problem is by using algebra. Here's a quick step-by-step guide:

Step 1: Write "A coffee and a muffin cost \$7.40 in total." Algebraically

$$c + m = 7.40$$

Step 2: Write "The muffin costs \$1 more than the coffee." Algebraically

$$m - c = 1$$

Step 3: Then, take the statement from Step 2 and rearrange for m

$$m = 1 + c$$

Step 4: Substitute the new equation for m into the statement in Step 1

$$c + (c + 1) = 7.40$$

Step 5: Rewrite the equation in Step 4 to make it simpler

$$2c + 1 = 7.40$$

$$2c = 7.40 - 1$$

$$2c = 6.40$$

Step 6: Solve for m!

$$c = 3.20$$

Therefore, the coffee is worth \$3.20!

Feedback (correct):

The question stated: A coffee and a muffin cost \$7.40 in total. The muffin costs \$1 more than the coffee. How much does the coffee cost?

It's often helpful to break these questions down into an algebraic format. Try and give it a go!

The question states "A coffee and a muffin cost \$7.40 in total."

How can we write this algebraically? Choose from the following.

Hint: Let c represent "coffee" and m represent "muffin"

a) $c + m = 7.40$

b) $c - m = 7.40$

c) $c + 7.40 = m$

The question also states "The muffin costs \$1 more than the coffee."

How can we write this algebraically? Choose from the following.

a) $m - c = 1$

b) $m + 1 = c$

c) $c + m = 1$

[Next page]

The question stated:

A coffee and a muffin cost \$7.40 in total. The muffin costs \$1 more than the coffee. How much does the coffee cost?

From the question (and our previous algebra), we know that:

1) $c + m = 7.40$

2) $m - c = 1$

Let's combine these statements to find c .

First, take statement (2) and rearrange for m . How can you rewrite (2) to put m on the left hand side? Choose from the following:

a) $m = 1 + c$

b) $m = 7.40 + 1$

c) $m = 1 - c$

Next, let's substitute our new equation for m into statement (1). Choose from the following:

a) $c + (1 + c) = 7.40$

b) $(1 + c) - c = 7.40$

c) $c = 1 + (c - 1)$

[Next page]

The question stated: A coffee and a muffin cost \$7.40 in total. The muffin costs \$1 more than the coffee. How much does the coffee cost?

You've worked out that:

$$c + (1 + c) = 7.40$$

Let's solve for c !

First, let's make that equation simpler. How can we rewrite this equation? Choose from the following:

a) $2c + 1 = 7.40$

$$2c = 6.40$$

b) $c + 1 = 7.40 + c$

$$c = 7.40 + c$$

c) $2c + 1 = 7.40$

$$2c = 5.40$$

Can you work out what the coffee is worth?

a) \$3.20

b) \$4.20

c) \$6.40

2. It takes 2 cinema attendants 2 hours to sell 200 tickets. How many hours would it take 6 attendants to sell 600 tickets?

Correct: 2 hours

Incorrect-1: 6 hours

Incorrect-2: 3 hours

Incorrect-3: 1 hour

Feedback:

The question stated: It takes 2 cinema attendants 2 hours to sell 200 tickets. How many hours would it take 6 attendants to sell 600 tickets? The correct answer is 2 hours. Each attendant can sell 100 tickets in ONE hour, there are 6 attendants, and so over 6 hours each of them sells 100 tickets. There 6 attendants so together they can sell 600 tickets.

3. An oak forest is suffering from an infectious, fast-spreading disease. Presently, the number of infected trees doubles every year. After 8 years, the forest has 5000 infected trees. After how many years would you predict the forest to have had 2500 infected trees?

Correct: 7 years

Incorrect-1: 4 years

Incorrect-2: 16 years

Incorrect-3: 2 years

Feedback:

The question stated: An oak forest is suffering from an infectious, fast-spreading disease. Presently, the number of infected trees doubles every year. After 8 years, the forest has 5000 infected trees. After how many years would you predict the forest to have had 2500 infected trees? The correct answer is 7 years. If, after 8 years, the forest has 5000 infected

trees, and the number of infected trees doubles in size every year, the forest will contain half the number of infected trees 1 year earlier. That is, after 7 years, we would expect there to have been 2500 infected trees.

4. A laptop with a warranty costs \$600 in total. The laptop costs \$500. How much does the warranty cost?

Correct: \$100

Incorrect-1: \$50

Incorrect-2: \$25

Incorrect-3: \$550

Feedback:

The question stated: A laptop with a warranty costs \$600 in total. The laptop costs \$500. How much does the warranty cost? The correct answer is \$100. You can work this out by subtracting the cost of the laptop from the total cost (i.e. \$600 - \$500).

5. It takes 1 chef 4 hours to make 4 cakes. How many hours would it take 1 chef to make 16 cakes?

Correct: 16 hours

Incorrect-1: 4 hours

Incorrect-2: 8 hours

Incorrect-3: 6 hours

Feedback:

The question stated: It takes 1 chef 4 hours to make 4 cakes. How many hours would it take 1 chef to make 16 cakes? The correct answer is 16 hours. If it takes 1 chef 4 hours to make 4 cakes, then we know it takes 1 chef 1 hour to make 1 cake. Knowing that, you can determine that 1 chef takes 16 hours to make 16 cakes.

6. A farmer has 20 cows. Every month the farmer buys 10 new cows. How many months would it take for the farmer to have 40 cows?

Correct: 2 months

Incorrect-1: 3 months

Incorrect-2: 4 months

Incorrect-3: 1 month

Feedback:

The question stated: A farmer has 20 cows. Every month the farmer buys 10 new cows.

How many months would it take for the farmer to have 40 cows? correct answer is 2 months. If the farmer buys 10 new cows every month, and he has started with 20 cows, it would take 2 months for the farmer to have 40 cows (i.e. $40 - 20 = 20$. $20/10 = 2$).

Test questions (lure):

1. A shirt and a jacket cost \$18 together in total. The shirt costs \$10 more than the jacket. How much does the jacket cost?

Correct: \$4

Incorrect-1: \$8

Incorrect-2: \$9

Incorrect-3: \$6

2. If it takes 4 dogs 4 minutes to eat 4 bowls of food. How many minutes would it take 3 dogs to eat 3 bowls?

Correct: 4 minutes

Incorrect-1: 3 minutes

Incorrect-2: 6 minutes

Incorrect-3: 2 minutes

3. In the city, there is a popular technology store. Every hour, a line of people waiting outside the store triples. After 3 hours, there are 90 people waiting in line. How many hours did it take for the line to reach 30 people?

Correct: 2 hours

Incorrect-1: 1 hour

Incorrect-2: 3 hours

Incorrect-3: 4 hours

Test questions (no lure):

4. A phone and a wallet cost \$1000 in total. The phone costs \$900. How much does the wallet cost?

Correct: \$100

Incorrect-1: \$50

Incorrect-2: \$200

Incorrect-3: \$900

5. A monkey can eat 6 bananas in 1 hour. How many hours would it take the monkey to eat 18 bananas?

Correct: 3 hours

Incorrect-1: 5 hours

Incorrect-2: 6 hours

Incorrect-3: 1 hour

6. An art gallery has 100 artworks. The number of artworks is expected to increase by 1 every week. How many weeks would it take the art gallery to have 105 artworks?

Correct: 5 weeks

Incorrect-1: 100 weeks

Incorrect-2: 10 weeks

Incorrect-3: 1 week

Visual Descriptions of Hypothesised Relationships

The following figures describe the expected pattern of results for each analysis conducted within Papers 1 to 4 that did not reach statistical significance. This is included to assist the reader in determining when the general pattern of results are consistent with the hypothesized pattern. The remaining hypotheses in the thesis reached statistical significance and are described in detail in the body of the thesis.

Paper 2

Study 1. Study 1 used a 4 (test point) x 2 (low load or high load) mixed design with N=107. The dependent variable differed for the two primary analysis. For the behavioural analysis, CRT performance was the dependent variable—the hypothesis for this analysis is presented in Figure 1. For the conflict and working memory engagement analysis, a binary (correct/incorrect) score was the dependent variable variable—the hypothesis for this analysis is presented in Figure 2.

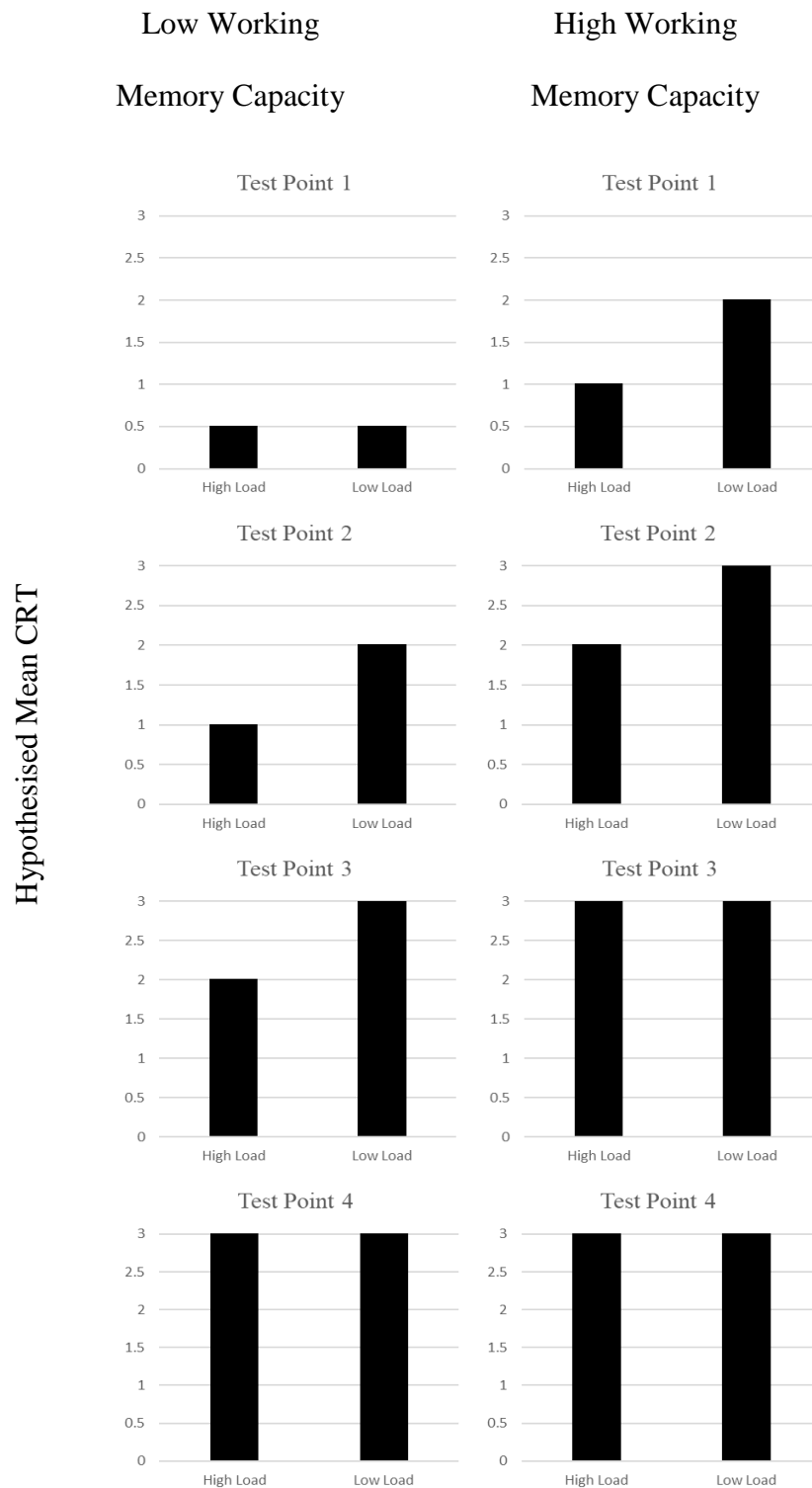


Figure 1. Graphical representation of the hypothesis for the relationship between DSE (test point), constraint (low or high load), and working memory capacity.

Low Conflict

High Conflict

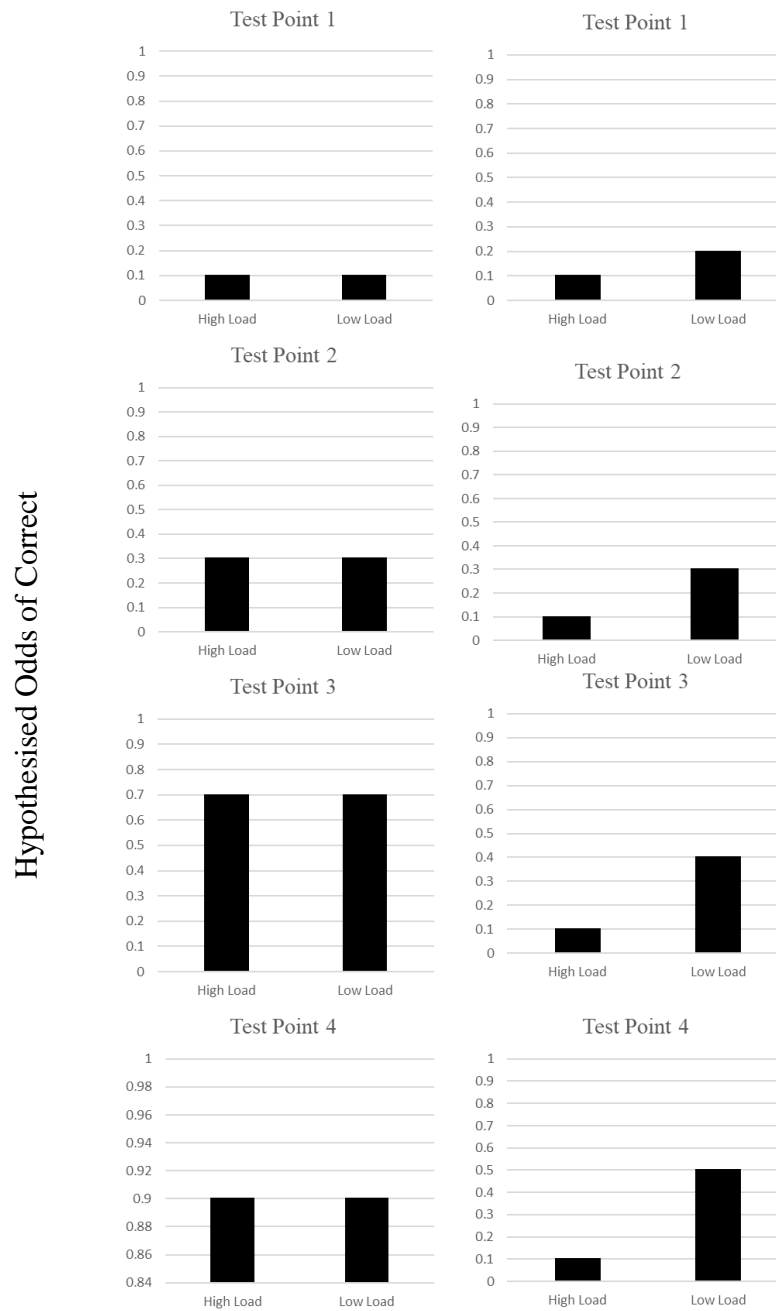


Figure 2. Graphical representation of the hypothesis for the relationship between conflict, DSE (test point) and constraint (load) on the odds of correct responding.

Study 2. Study 2 used a 4 (test point) x 2 (response) within-subject design with N=125. The hypothesised relationship between test point and response is depicted in Figure 3.

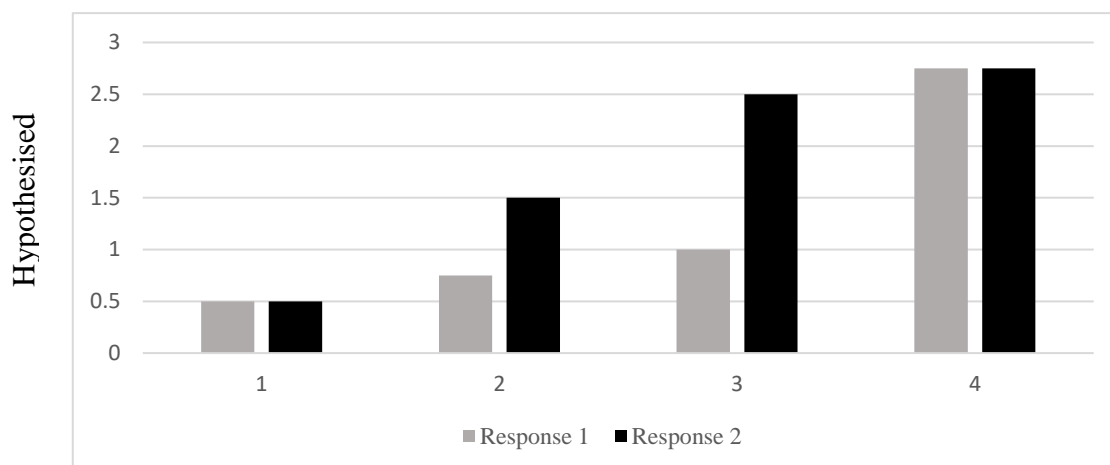


Figure 3. The hypothesised relationship between test point and response on performance on the conflict items.

Paper 3

A 4 (test point: T1, T2, T3, T4) x 2 (problem type: lure and no lure) x 2 (load condition: low and high load) mixed design was used.

Behavioural analysis. We examined the interaction between test point, load condition and WMC on performance on the lure items. The hypothesised relationship between these factors is identical to that in Paper 2 (see Figure 1.)

Strategy measured by fixations and dwell. To examine the strategies employed on lure items for correct- versus heuristic-respondents, a linear mixed model was run. The model included four predictors: item (1,2,3), test point (T1, T2, T3, T4), accuracy (correct, heuristic), and AOI (Selected, Other-Relevant, Other-1, Other-2). The dependent variables were the proportion of fixations and the proportion of dwell. The hypothesised relationship between these factors is depicted in Figure 4.

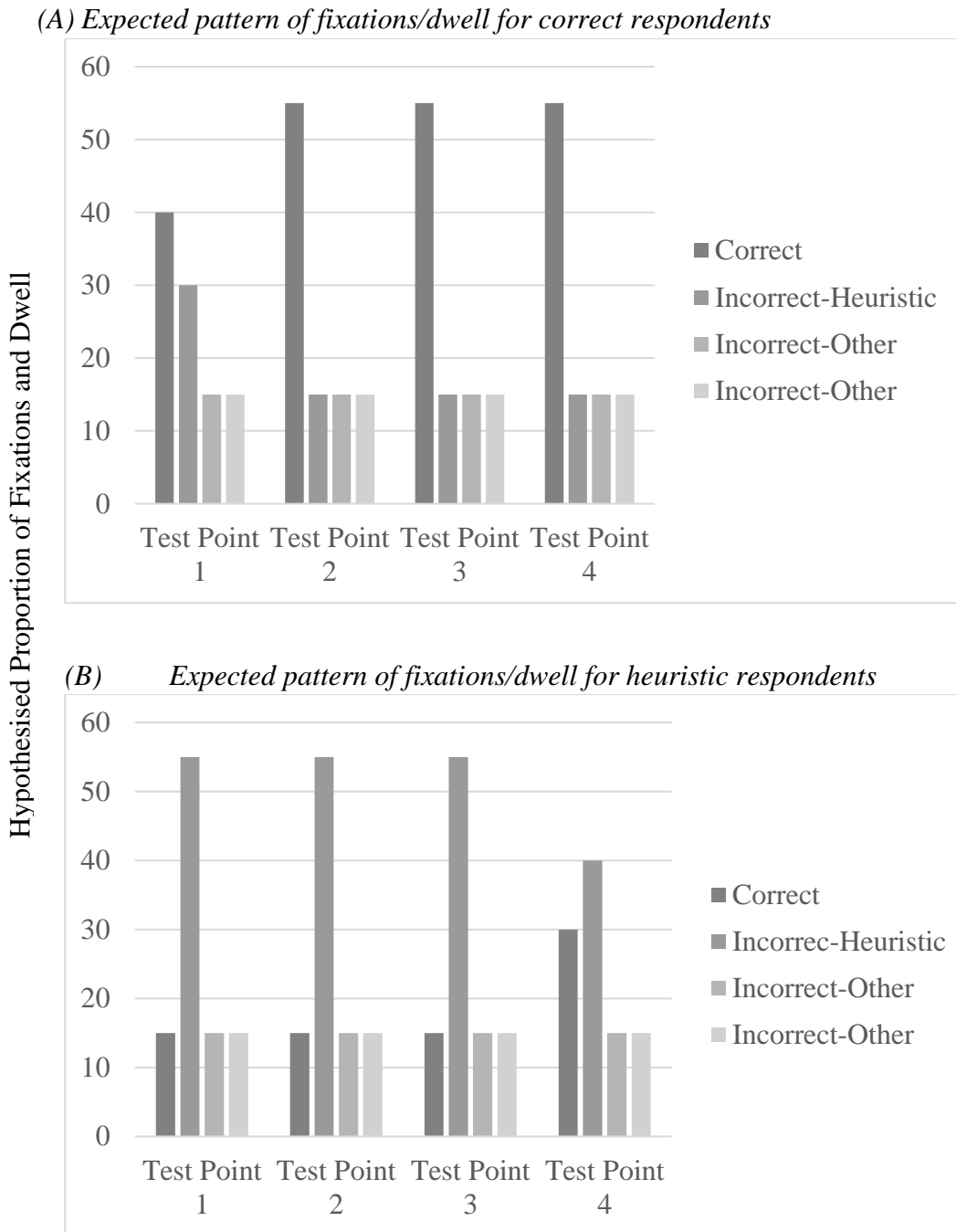


Figure 4. The hypothesised relationship between accuracy, test point and AOI (item was included as a covariate) on the proportion of fixations and dwell by test point and AOI.

Implicit conflict measured via the proportion of fixations and dwell. A linear mixed model was used to examine implicit conflict as measured by the proportion of fixations that occurred on each multiple-choice option. As above, AOIs were coded to reflect the ‘Selected’, ‘Other-Relevant’, ‘Other-1’ and ‘Other-2’ response options. The predictors were item (1,2,3: covariate only), test point (T1, T2, T3, T4), trial type (heuristic-lure,

correct-no lure), and AOI. The dependent variable was the proportion of fixations or the proportion of dwell. The hypothesised relationship between these factors is depicted in Figure 5.

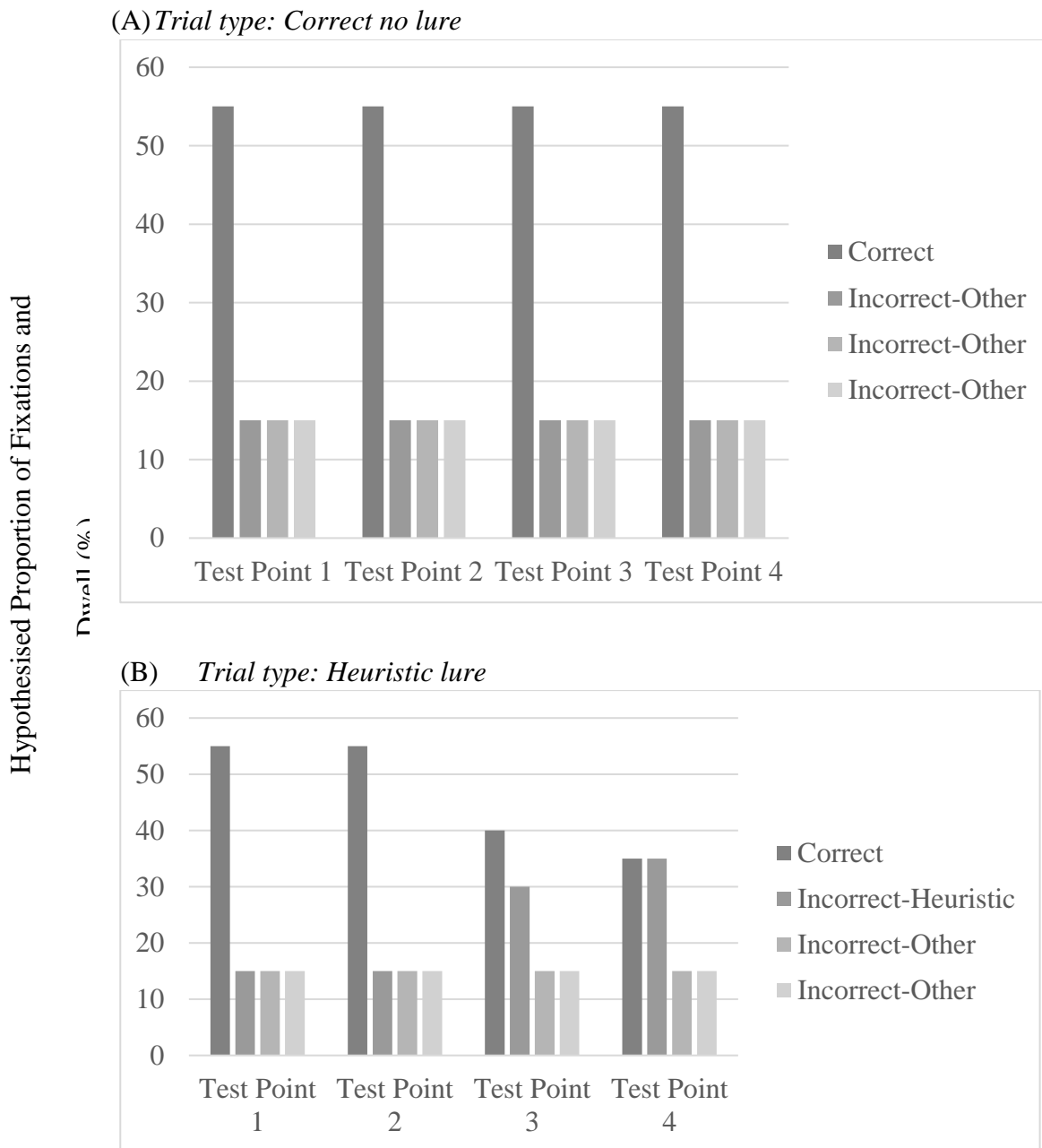


Figure 5. The hypothesised relationship between trial type, test point and AOI (item was included as a covariate) on the proportion of fixations and dwell by test point and AOI.

Paper 4

Implicit conflict. We conducted a two-step examination of the pattern of implicit conflict across test point (T1, T2, T3, T4) and trial type (CNL, HL). First, we used a single factor measure of implicit conflict, and second, we examined the spread of fixations across the four multiple-choice options. The hypothesis pertaining to the single factor measure is

depicted in Figure 6. The hypothesis pertaining to spread of fixations is identical to that portrayed in Figure 5 above.

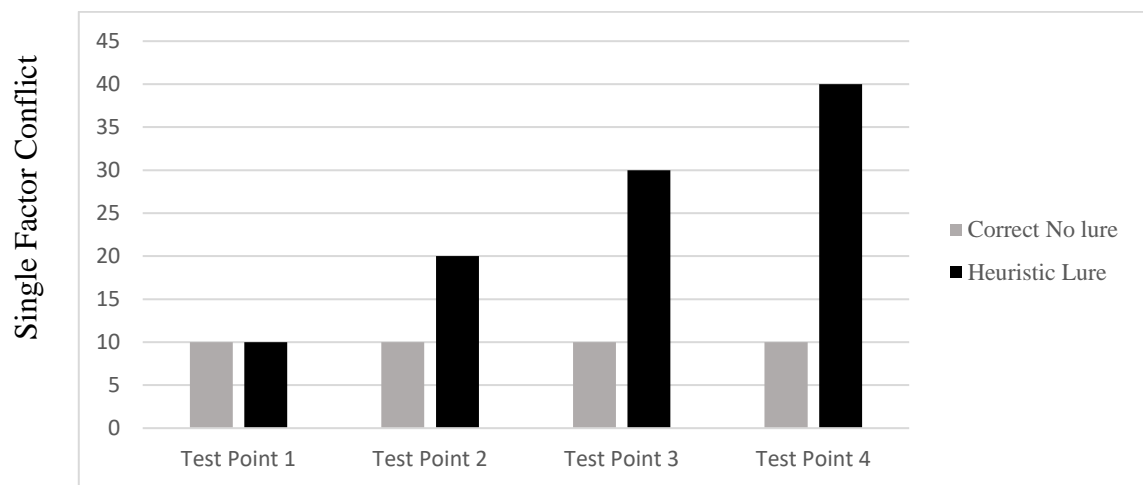


Figure 6. The hypothesised relationship between test point, trial type (correct no lure, heuristic lure) on the single factor conflict measure.