On Spoken Confidence: Characteristics of Explicit Metacognition in Reasoning

by

Kaiden M Stewart

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Psychology

Waterloo, Ontario, Canada, 2025

© Kaiden M Stewart 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Dr. Chris Fiacconi Associate Professor, Department of Psychology University of Guelph
Supervisor	Dr. Jonathan Fugelsang Professor, Department of Psychology University of Waterloo
Internal Members	Dr. Derek Koehler Professor, Department of Psychology University of Waterloo
	Dr. Clara Colombatto Assistant Professor, Department of Psychology University of Waterloo
Internal-external Member	Dr. Philip Curry Associate Professor, Department of Economics University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This research was conducted at the University of Waterloo by Kaiden Stewart under the supervision of Dr. Jonathan Fugelsang. As is customary, I use first-person singular language (i.e., "I") throughout, but the work contained in this document is coloured by the minds of several. All experiments herein were overseen by Dr. Fugelsang, and Experiments 1a and 2a were also overseen by Dr. Evan Risko.

All of the work presented in this thesis, unless otherwise attributed, has been conceptualized, designed, and executed primarily by Kaiden Stewart.

Abstract

In this thesis, I assess how explicit, subjective evaluations of confidence influence monitoring and control (i.e., metacognitive) processes in reasoning. Metacognitive processes play a crucial role in modern dual-process theories of reasoning and decision-making, the consequences of which have been implicated in numerous significant real-world decisional outcomes. It is tacitly assumed that monitoring one's reasoning for the purpose of optimal deployment of controlled, deliberative processing functions similarly to monitoring one's reasoning for the purpose of providing a judgment of confidence, despite evidence from other domains indicating otherwise. This thesis takes a critical step toward evaluating metacognitive theories of reasoning and their broader application by assessing the degree to which standard approaches represent realistic accounts of metacognitive processes. To aid in interpretation of the work directly testing this possibility, I first present six experiments addressing foundational issues with respect to the operation of metacognition in reasoning. Chapter 2 provides evidence for a causal relationship between confidence judgments and controlled behavior (specifically deliberation), a reality often assumed in the absence of direct evidence. I demonstrate across four experiments that processing manipulations affect confidence and influence control behavior, consistent with a causal relationship, but also that it is possible to target control behaviour without mirroring effects on confidence. Chapter 3 develops a simple predictive model of confidence that identifies heretofore unidentified, item-based predictors of confidence. This simple model allows a unique approach to testing the central question in Chapter 4. Chapter 4 investigates whether the relationship between confidence and controlled behavior partly depends on the requirement to make explicit confidence judgments. Using a paradigm adapted from research involving nonhuman primates, I compare implicit and explicit confidence conditions. Results reveal small differences in controlled behavior and substantial differences in monitoring.

In the present thesis, I provide evidence of plausibly systematic influences of common measurement approaches on reasoning. To this effect, it is likely that the reasoning processes in which individuals engage in day-to-day life are reliably different than those commonly assessed in the lab. This has practical, but also theoretical implications which I discuss.

Acknowledgements

There are several people to whom I am in debt, but foremost among them is my supervisor, Dr. Jonathan Fugelsang, without whom this journey would have been far less fruitful and enjoyable. For providing the unending positivity, fostering curiosity, and shouldering both of our shares of the anxiety, we'll never be even. For feedback and support, impromptu discussions, role modelling, and challenging, I would like to thank Drs. Derek Koehler, Clara Colombatto, Sam Johnson, Colin McLeod, Derek Besner, Evan Risko, and the rest of the faculty and staff at the University of Waterloo Department of Psychology.

I would like to thank the many members of the Reasoning and Decision-Making Lab with whom I crossed paths. To my colleagues: the fellow graduate students, postdocs, undergraduates, research assistants, I consider myself very lucky to have passed through this place at the same time as you. The camaraderie, engaging discussion, and general atmosphere will be missed (in fact, it already is). I would gladly study again through a once-in-a-generation global pandemic for the opportunity to share it with you all, though I could do without the pandemic part. To wit, I would specifically like *not* to thank COVID-19.

To my family: I feel your support always and your expectations never. All my accomplishments are yours as well. It is because of you that I have lofty goals and the belief that I can reach them. Thanks Mom and Dad, I love you to the moon and back. And to my wife, Kelsey: I could not, and would not, have done this without you. I cannot wait to see where life takes us, and I am grateful to have shared this place and these experiences with you. Let's roll.

Dedication

To my wife, Kelsey, without whom I would probably not have bothered with all this.

You’re simply the best

Table of Contents

Examining Committee Membership ii
Author's Declaration iii
Statement of Contributions iv
Abstract v
Acknowledgements vii
Dedication viii
List of Figures xii
List of Tables xiii
Quote xiv
Chapter 1: General Introduction 1
1.1 An Introduction to Metacognition 1
1.2 Characteristics of Metacognition 6
1.2.1 Characteristics of Monitoring 6
1.2.2 Characteristics of Control 9
1.3 Metacognition in Reasoning and Decision-Making 10
1.4 The Present Investigation 14
1.4.1 Transparency and Openness 17
Chapter 2: Does (Under) Confidence *Cause* Deliberation? 19
2.1 Introduction 19
2.2 Experiment 1a 22
2.2.1 Method 23
2.2.2 Results 25
2.2.3 Discussion 29
2.3 Experiment 1b 30
2.3.1 Method 30
2.3.2 Results 32
2.3.3 Discussion 40
2.4 Experiment 2a 42
2.4.1 Method 43
2.4.2 Results 45
2.4.3 Discussion 49

2.5 Experiment 2b 50
2.5.1 Method 50
2.5.2 Results 52
2.5.3 Discussion 55
2.6 Chapter Review and Discussion 56
Chapter 3: Refining an Item-Based Model of Confidence in Reasoning 59
3.1 Introduction 59
3.2 Experiment 3 62
3.2.1 Method 63
3.2.2 Results 65
3.2.3 Discussion 71
3.3 Experiment 4 72
3.3.1 Method 73
3.3.2 Results 74
3.3.3 Discussion 78
3.4 Chapter Review and Discussion 78
Chapter 4: On the Effects of Explicit Judgments of Confidence 82
4.1 Introduction 82
4.2 Experiment 5 83
4.2.1 Method 83
4.2.2 Results 86
4.2.3 Discussion 95
4.3 Experiment 6 97
4.3.1 Method 98
4.3.2 Results 100
4.3.3 Discussion 105
4.4 Chapter Review and Discussion 108
Chapter 5: General Discussion 112
5.1 Thesis Summary 112
5.2 Situating these Results within the Meta-reasoning Framework 115
5.3 Evidence Accumulation and Dual Process Theories 118
5.4 Limitations of Evidence and Generalizability 119

5.5 Mechanistic Considerations and Future Directions 121
5.6 Conclusions 122
References 124
Appendix A: Items Used in All Experiments 154
Appendix B: Seven-point FOR Judgment 160
Appendix C: Six-point FOR Judgment 161

List of Figures

Figure 1. An illustration of monitoring from De Neys (2023) 13
Figure 2. Feeling of Rightness by Delay condition in Experiment 1a 27
Figure 3. Rethinking Time by Delay condition in Experiment 1a..... 27
Figure 4. Answer Change by Delay condition in Experiment 1a 28
Figure 5. Feeling of Rightness by Delay condition in Experiment 1b 34
Figure 6. Rethinking Time by Delay condition in Experiment 1b 35
Figure 7. Answer Change by Delay condition in Experiment 1b..... 35
Figure 8. Feeling of Rightness by Delay condition in Experiments 1a and 1b 38
Figure 9. Rethinking Time by Delay condition in Experiments 1a and 1b 39
Figure 10. Answer Change by Delay condition in Experiments 1a and 1b..... 39
Figure 11. Feeling of Rightness by Delay condition in Experiment 2a..... 47
Figure 12. Answer Change by Delay condition in Experiment 2a 48
Figure 13. Feeling of Rightness (FOR) by Delay condition in Experiment 2b..... 54
Figure 14. Answer Change by Delay condition in Experiment 2b..... 54
Figure 15. Rethinking Time by Condition and Confidence in Experiment 5..... 89
Figure 16. Answer Change by Condition and Confidence in Experiment 5 90
Figure 17. Predicted probabilities of a high confidence response by Condition and Accuracy in Experiment 5..... 95
Figure 18. Predicted values of Rethinking Time by Condition and Confidence in Experiment 6..... 102
Figure 19. Predicted Probabilities of Answer Change by Condition and Confidence in Experiment 6..... 102
Figure 20. Predicted Confidence Judgments by Condition and Response Time in Experiment 6..... 104
Figure 21. Predicted Confidence Judgment by Condition and Accuracy in Experiment 6..... 105

List of Tables

Table 1. Confidence predicted by item characteristics, Experiment 3 69
Table 2. Confidence predicted by item characteristics, Experiment 4 78

Quote

We have to remember that what we observe is not nature in itself,

but nature exposed to our method of questioning

- Werner Heisenberg

Chapter 1: General Introduction

Imagine your son sends you a copy of *Thinking, Fast and Slow* by Daniel Kahneman (2011) to help explain what he is studying in graduate school. You read it, and it is full of novel, unfamiliar information. Perhaps the concepts and theories introduced make sense and you are immediately able to integrate them with information you already have; perhaps they do not, and you cannot. But how do you *know*? And if you cannot immediately make sense of this information, what do you *do* about it? Humans have a unique ability to represent our own cognitive processes and alter them adaptively to operate effectively within our environment (Carruthers & Williams, 2019; Carruthers & Williams, 2022; Fleming & Lau, 2014). It is this ability that allows us to effectively engage in the breadth of cognitive activities that colour human life (Fleming, 2024; Rouault et al., 2018). It has even been suggested that representation of mental states underlies or is to varying degrees related to the very nature of conscious experience itself (e.g., Koriat, 2007; Lau, 2019; Nelson, 1996; Snodgrass et al., 2009). Despite the wide-ranging scope of the characterization and implications of the human ability to represent and adapt one's cognitive processes, the field has been at least partially unified under the umbrella of "metacognition."

1.1 An Introduction to Metacognition

Generally speaking, metacognition refers to any instance where one represents their own cognitive process in some way, usually for the purpose of some form of top-down adaptation of those processes toward the successful completion of some task. To that end, it has often been described as "cognition about cognition", or perhaps more colloquially as "thinking about thinking" (Flavell, 1979; Nelson & Narens, 1994). Within this domain, there is normally a distinction between two distinct types of metacognitive processes: individuals evaluate their own

cognitions, a family of processes labelled "monitoring" (for a broad overview, see Schraw, 2009), which allows them to strategically and adaptively alter or differentially engage their own cognitive operations, a family of processes labelled "control" (Dunlosky & Metcalfe, 2009; Koriat, 2007). A litany of work has illuminated the characteristics and functions of each branch of metacognition across various domains of human cognition, including perceptual decision-making (e.g., Fleming, Dolan, et al., 2012; Pescetelli et al., 2016, Shekhar & Rahnev, 2021, for a review, see Rahnev, 2021), episodic (e.g., Perrin & Sant'Anna, 2022; Whatley & Castel, 2022), semantic (e.g., Burratti et al., 2013; Koriat & Goldsmith, 1996), and working memory (e.g., Adam & Vogel, 2016; Cowan et al., 2016; Rademaker et al., 2012), as well as higher-level reasoning and decision-making (e.g., Prowse Turner & Thompson, 2009; Shynkaruk & Thompson, 2006; Thompson et al., 2011; for theoretical reviews, see De Neys 2017, and Ackerman & Thompson, 2017). In all, an explosion of research has investigated monitoring and control processes across the range of human cognition.

Recent work has undertaken the task of exploring and describing the computational functions (e.g., Kepecs & Mainen, 2012) and/or neurological hardware (e.g., Insabato et al., 2010; Timmermans et al., 2012) that provide for the richness and adaptability of human metacognition. This work has yielded numerous theoretical advances in the understanding of the nature and scope of metacognitive processes. A useful framework for understanding metacognition with respect to the features of mental representation is provided by Fleming and colleagues (2012). They describe dimensions along which representations vary, toward a taxonomy of what is and is not metacognitive. First-order behaviours, for example, are contingent only on the processing of some stimulus; second-order behaviours are contingent on first-order behaviours, and so on. Conversely, representations can be at the object-level or they

can be meta-level. Object-level representations are of external stimuli; meta-level representations are of mental states arising from the processing of external stimuli. One way individuals can engage in metacognitive processes is to directly represent the output of some prior cognitive process; this is, by definition, metarepresentational in nature. However, meta-level representation is not necessarily required in order to adapt one's behaviour contingent on outputs from past behaviour (Fleming, Huijgen, & Dolan, 2012).

Illustratively, let us briefly consider a disjunction between common models of metacognitive monitoring. Note that the purpose of this thesis is not to adjudicate between these models, and as such this will not be near an exhaustive review. One common historical distinction is between inferential and direct access models of metacognition. In inferential models, metacognitive judgments are informed by information external to the first-order judgment. A great example of this type of model, the Self-Consistency Model, is introduced by Koriat (1993). In this model, metacognitive judgments are divorced from some characteristics of performance on the object-level task. Subjective metacognitive reports are informed by more than the object-level information; they involve an averaging of task-related and experiential information into a unified and usable representation of performance. That is, while subjective reports require one's own processing to be representable, they do not necessarily require that representation to be only of task-related behaviour. Conversely, no such provision is allowed in direct access models. In direct access models, metacognitive judgments are thought to be translations of task-related cognition; a high degree of evidence in favour of one option over the other, for example, predicts *both* the choice of that option and confidence in one's choice. In direct access models, metacognitive monitoring can be achieved without meta-level representations. In both types of models, metacognition is higher-order to cognition (Fleming &

Daw, 2017; Overgaard & Sandberg, 2012), and different methods of elicitation correspond to different assumptions about the nature of metacognition (Kornell, 2014).

In conjunction with the main focus of the latter parts of this thesis, let us focus on the various ways researchers can elicit confidence. Confidence has been operationalized numerous ways, but in the interest of breadth and for the purposes of this thesis, confidence will refer to any judgment or behaviour corresponding to the assessment of one's behaviour (Fleming, 2024; Fleming & Lau, 2014; Kleitman & Stankov, 2007; Yeung & Summerfield, 2012), past (retrospective confidence; Dougherty et al., 2005; Hines et al., 2009) or future (prospective confidence; Castel et al., 2007; Koriat & Bjork, 2005, 2006; McCabe & Soderstrom, 2011; Rhodes & Castel, 2008). The simplest method of assessing individuals' confidence in their performance on some task is to simply ask them to rate numerically on some scale how likely they are to have responded correctly. Hereon, I will refer to these as explicit or subjective confidence judgments, and I will collapse under this umbrella a number of judgment types that involve explicit report of the estimated success of past or future cognition (Feelings of Rightness [FORs; Thompson et al., 2011; Wang & Thompson, 2019], Feelings of Knowing [FOKs; Metcalfe, 1986; Schacter, 1983], Judgments of Learning [JOLs; for a review, see Rhodes, 2016], etc.). These measures are by far the most popular cross-domain in metacognition research. In addition to explicit confidence judgments, there are a number of behavioural measures that have been used to study confidence in humans and animals. For example, researchers have added a response option for participants to indicate they are uncertain (Smith et al., 2006; Watson et al., 1973), added an option where participants can choose to withhold a response (Arnold & Prike, 2015; Sidi & Ackerman, 2024; Undorf et al., 2021), or elicited post-decision wagers whereby participants must evaluate their performance in order to optimize the payout they receive on the

task (Fleming & Dolan, 2010; Schurger & Sher, 2008). Hereon, despite the occlusion of meaningful differences within these categories, I will often refer to these as *explicit* (confidence judgments) and *implicit* (evaluated by behavioural measures in the absence of an explicit confidence judgment), for the sake of simplicity.

Because these behavioural measures escape the need for subjects to respond to a verbal prompt, they have also been used to study metacognition in nonhuman animals. This has raised the question of whether they are ultimately measuring the same construct as explicit confidence responses (Kepecs & Mainen, 2012). For example, Proust (2012) draws a distinction between private metacognition, wherein individuals adapt their behaviour based on internal cues, and mindreading, which requires meta-level representations of internal mental states. Only the former has been demonstrated in animals and young children, who do not have the capacity to represent minds, others' or their own (Proust, 2012; Hampton, 2009). Likewise, Shea and colleagues (2014) use systems language to differentiate between metacognition destined for internal use (System 1) and for external broadcast (System 2). Elsewhere, researchers have argued that subjective reports of confidence can colour confidence judgments with task-unrelated factors (Castel et al., 2007; Koriat, 1993) or correspond to different underlying object-related representations than behavioural or neuroscientific measures (Bang et al., 2020). So, while behavioural adaptations of confidence judgments allow for the study of metacognition in those who may be unable to verbally report their confidence, their use raises a heretofore underexplored question: to what extent do they differ mechanically from subjective, verbal reports of confidence? Or perhaps more aptly, to what extent do in-laboratory investigations using subjective reports of confidence alter the metacognitive calculus beyond what individuals might use in their everyday lives to monitor and control their cognitive behaviour?

1.2 Characteristics of Metacognition

Recent years have seen an explosion of research investigating the characteristics of and the mechanisms behind metacognitive processes. Perhaps because of the amenability of memory research to laboratory settings, a disproportionate amount of metacognitive research has been related to memory. Though far from the only subdomain in which this research occurs, "metamemory" research has been largely responsible for the bulk of the empirical advances in our collective understanding of metacognition.

1.2.1 Characteristics of Monitoring

Borne partially of the aforementioned tension between direct access and inferential models of metacognition, metacognitive monitoring has become widely considered a process of cue utilization (Koriat, 1997). The cue utilization approach allows for far greater variability in subjective reports of confidence than does the direct access view, as it allows for multi-source cues to collaboratively influence metacognitions (Jang & Nelson, 2005; Mueller & Dunlosky, 2017; Tauber & Rhodes, 2012). However, this approach canonically segregates cues by their relation to the target task, and different cue utilization accounts claim different weights vis-à-vis confidence. Koriat (1997) originally separated *intrinsic* cues from *extrinsic* cues, where intrinsic cues are referent to the characteristics of items themselves, and extrinsic cues are referent to some other aspect of processing unrelated to the nature of to-be-processed items. For example, word frequency (Benjamin, 2003) and concreteness (Hertzog et al., 2003; Tullis & Benjamin, 2012; Witherby & Tauber, 2017) are stimulus-intrinsic factors that reliably impact JOLs, although not always in directional concordance with the effects they have on memory performance (See Fiacconi & Dollois, 2012 for a review).

It was originally argued that confidence judgments (specifically JOLs) placed heavy weight on intrinsic cues and less or no weight on extrinsic cues, suggesting some degree of discrimination between the two in the metacognitive system of participants (Koriat, 1997), for which there is some evidence. For example, Koriat and colleagues (2004) demonstrated that individuals' subjective estimates of performance were far less sensitive to an increased time interval between study and test than was their performance. There is a family of evidence in support of this claim about the insensitivity of metacognition to extrinsic cues (e.g. Castel, 2008; Kornell & Bjork, 2009). However, a family of research has demonstrated that subjective evaluations of confidence are in some instances oversensitive to extrinsic factors, even those that do not impact actual performance. For example, Rhodes and Castel (2008) demonstrated that participants predict better memory performance for words presented in larger font, despite little impact of font size on performance.

Numerous attempts have been made to identify underlying characteristics that predict when and how metacognitive cues will colour judgments made by participants. A number of researchers claim that at least some of these effects are due to the beliefs individuals hold about their own cognitions. Mueller and colleagues (2014), for example, demonstrated that the font size effect was almost entirely driven by beliefs participants had about font size and performance in the task. Similarly, individuals' metacognitive beliefs have been implicated in other instances of extrinsic cue effects on JOLs (Mueller et al., 2013; Mueller et al., 2014; Mueller et al., 2016). A counter-explanation of these types of effects has been made on the basis of ease of processing (termed fluency). In essence, some have argued that at least some extrinsic factors make stimuli easier or harder to process, and thus despite irrelevance of these cues vis-à-vis the internal content of stimuli, individuals base metacognitive judgments on the resulting ease or difficulty of

processing (Fiacconi et al., 2020). With respect to fluency, it has been demonstrated that metacognitive judgments are influenced by ease of processing at every stage of common laboratory tasks, from perception of stimuli (Besken & Mulligan, 2013, 2014; Murphy et al., 2022; Reber & Schwarz, 1999; Susser et al., 2013; Undorf et al., 2017), to their encoding (Hertzog et al., 2003; Koriat & Ma'ayan, 2005) and retrieval (Benjamin et al., 1998; Koriat & Ma'ayan, 2005; Matvey et al., 2001). To add a layer of complexity, it has also been demonstrated that individuals' beliefs *about* processing ease moderate effects of ease itself (Hertwig et al., 2008; Miele et al., 2011; Miele et al., 2013).

In sum, current understanding of metacognitive monitoring implicates a complex, multi-cue process with unclear relations between the numerous factors that influence metacognitive judgments. Indeed, much of the current work aims to understand and characterize the contributions of single cues on confidence, and the field moves forward piecemeal. However, there are researchers advancing knowledge of the interplay between multiple cues: Bröder and Undorf (2019) and Undorf and colleagues (2018) demonstrate that participants combine multiple manipulated extrinsic cues in the formation of JOLs. Another interesting approach to the study of multiple cues is derived from Koriat's (2012) Self-Consistency Model introduced earlier, wherein it is posited that confidence judgments are the product of the general alignment of the totality of cues indicating a particular response. If there is general agreement between the cues for a given item, it will be the case that i) most participants will provide the same response, and ii) confidence will generally be high for that response. A number of researchers have demonstrated such patterns in various cognitive tasks (Boundy-Singer et al., 2023; Brewer & Sampaio, 2006, 2012; Caziot & Mamassian, 2021).

1.2.2 Characteristics of Control

Modern research has also described characteristics of control processes. For the purpose of the present thesis, I will focus mainly on two instances of controlled behaviour in investigations of memory and perception. First, does controlled behaviour, defined specifically to the subdomain, arise from metacognitive estimates of performance? It seems so: in memory research, it has been demonstrated that individuals, when given the opportunity, will allocate their study time based on subjective confidence judgments about their performance (Bjork et al., 2013; Metcalfe & Finn, 2008; Son & Metcalfe, 2000; for a review, see Finley et al., 2010; and Metcalfe, 2009). While this generally materializes as participants engaging in more controlled processing of particularly uncertain items, it should be noted that task demands sometimes erase or even reverse this pattern (see Metcalfe & Kornell, 2005; Thiede & Dunlosky, 1999). Even in these apparent counterexamples, however, it is acknowledged that monitoring processes are a primary determinant of controlled behaviour. Similarly, it has been demonstrated that subjective confidence judgments drive controlled behaviour in perceptual decision-making tasks (Desender et al., 2018).

A second ostensible instance of controlled behaviour appears in the work on JOL reactivity. Several researchers have demonstrated that the act of making JOLs in and of itself alters ultimate memory performance (see Double & Birney, 2019; Undorf et al., 2024 for reviews). In the classical example, memory performance in the JOL condition exceeds performance in the non-JOL condition (Dougherty et al., 2005; Janes et al., 2018; Soderstrom et al., 2015), suggesting some processing difference by virtue of the requirement to assess one's metacognition. Identifying a mechanism by which JOLs affect cognition has proven difficult, as reactivity is often absent (Kelemen & Weaver, 1997; Benjamin et al., 1998; Tauber & Rhodes,

2012), and occasionally in the opposite direction, where JOLs *hinder* memory performance (Mitchum et al., 2016; Rivers et al., 2021). However, two primary proposed mechanisms persist. In the Strengthening account (Soderstrom et al., 2015), metacognitive monitoring involves evaluation of memory cues, and the evaluation of these cues in and of itself strengthens their link to the response and improves performance. In the Changed-goal account (Mitchum et al., 2016), JOLs serve as a cue to individuals where best to direct their effort; boosted performance is the result of this targeted control behaviour. While the definitive mechanism behind JOL reactivity has not been heretofore arbitrable, some evidence in favour of the Changed-goal account has mirrored the discussion in the previous section about the effects of explicit, subjective confidence judgments. Across three experiments, Myers (2024) varies the characteristics of JOLs (duration, granularity, and required effort) that, under a strengthening account, should increase the extent to which performance is reactive to them, and finds no such pattern. This is taken as support of a control account of reactivity, but also as a suggestion that the characteristics of explicit confidence judgments is largely inconsequential; what matters vis-à-vis JOL reactivity is that these judgments are made at all.

1.3 Metacognition in Reasoning and Decision-Making

Understanding of metacognitive processes underlying decision-making is partially derived from understanding of metacognitive processes in other domains, some of which are elucidated above. While it is often argued that claims about metacognition in any given subdomain are domain general (McCurdy et al., 2013; Song et al., 2011, but see Fitzgerald et al., 2017; Fleming et al., 2010) it is critical to understand the decision-making context in which metacognition operates for full comprehension. Decision-making has long been construed as arising from two distinct types of cognitive processes (De Neys, 2006a; Evans, 2008; 2010;

Evans & Stanovich, 2013; Frankish & Evans, 2009; Sloman, 1996). While the exact characteristics of each type has been a topic of debate (e.g., Keren & Schul, 2009; Melnikoff & Bargh, 2018), it is generally agreed that Type I processes are autonomous and do not require resources from working memory, and Type II processes require working memory resources and are required to solve problems that require abstraction or mental simulation (Evans & Stanovich, 2013; Stanovich, 2012).

Because of this argued qualitative distinction between Types I and II, numerous example tasks have been studied which allegedly *require* Type II processing in order to solve correctly. For instance, take the following deductive reasoning syllogism:

No animals are dogs.

Some lions are dogs.

Therefore, some lions are not animals.

Conventionally, correct responding to this type of problem requires a participant to evaluate the logic validity of the statement, which in this case is valid: if one takes the premises as true, the conclusion must also be true. However, some information is provided that is *unrelated* to the logical structure, namely the individual's belief-based assessment of the semantic content of the syllogism, which suggests that the syllogism is untrue (i.e., the opposite response). The traditional dual-process account of individuals' attempts to answer this type of problem involves Type I processes favouring the belief-based response (i.e., invalid), while Type II processes favour the logical response (i.e., valid; Evans et al., 1983; Evans et al., 2001). One account is that Type I processes complete faster and without the need for working memory resources, and thus belief-based responses will increase under time pressure (Evans & Curtis-Holmes, 2005) or

under cognitive load (De Neys, 2006b). Under this account, Type II processing was long held as superior, as it allowed reasoners to be spared from a number of reasoning failures and biases thought to be the result of the failure to engage deliberative, Type II processing (Evans, 2010). However, it became clear that this distinction was due, at least in part, to the fact that common laboratory tasks had been generated to create exactly this scenario, and as such, originalist conceptions of Type II processes as normative were probably fallacious (Evans, 2019). However, a renewed focus on the value of deliberation has arisen more recently. Lack of deliberation is implicated in a number of modern, real-world failures of reasoning, including epistemically suspect beliefs (Lobato et al., 2014), receptivity to bullshit (Pennycook, Cheyne, et al., 2015), and acceptance of misinformation and fake news (Pennycook & Rand, 2019; for a review, see Pennycook, 2023).

With renewed focus on the value of analytic of deliberative processing comes a renewed focus on predictors of the engagement of such processing. Given the evidenced benefits of analytical processing on consequential, real-world decisions, it is perhaps the most critical endeavour in the field to understand the circumstances in which it is (and is not) activated. Correspondingly, much attention has been given to characteristics of reasoners (Frederick, 2005; Littrell et al., 2020) and reasoning problems (De Neys & Glumicic, 2008) that predict deliberation, but modern theorizing increasingly includes metacognition as a driving factor in the switch from intuitive to deliberative processing. Broad theories by Thompson and colleagues (2011), Pennycook, Fugelsang, and Koehler (2015) and De Neys (2023) explicitly incorporate monitoring as part of the switch function (see also Ackerman & Thompson, 2017). Interestingly, all conceive of deliberative, traditionally Type II processing as essentially a control behaviour, activated by the output of a monitoring system. Under this reality, understanding of meta-

reasoning systems might be a key to widespread improvement of human decision-making. However, there are fine differences in the articulated monitoring processes involved in each model. For example, Thompson and colleagues (2011) propose that Type I, intuitive outputs are accompanied by a phenomenological Feeling of Rightness which arises from the ease or difficulty with which an answer sprung to mind. Similarly, Pennycook, Fugelsang, and Koehler's model (2015) is generally agnostic vis-à-vis the exact mechanism monitoring but is consistent with the possibility that a phenomenological sense of confidence probably arises from the asymmetry between multiple Type I inputs. Perhaps the most thoroughly-articulated account of monitoring processes in reasoning comes from De Neys (2023). In this proposed model, multiple competing intuitions are produced by an intuitive System I (see Figure 1), another function of which is uncertainty monitoring. If there is a large difference in the evidence for two alternative options, a sense of certainty will be produced (an implication of which is that System II will not be engaged). If, however, the relative evidence is approximately equal, System I will output a sense of uncertainty.

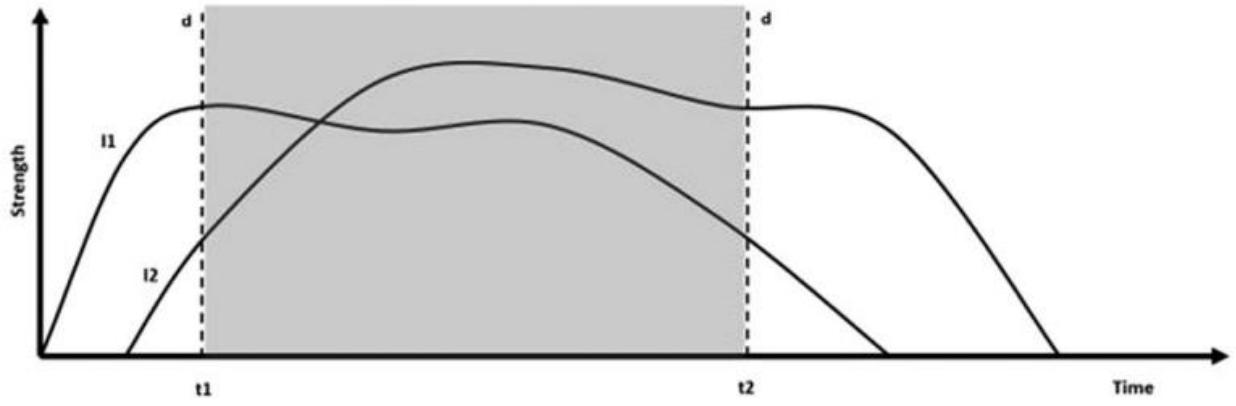


Figure 1. An illustration of monitoring from De Neys (2023). Represented are relative evidence weights for two competing intuitions (I1 and I2). Evidence initially accumulates in favour of I1 (the traditionally intuitive response), but an uncertainty monitoring system activates deliberation which in turn accumulates sufficient evidence to supplant that response with I2 (the traditionally deliberative response).

There are some natural questions to be raised. For example, are these accounts reconcilable? The answer seems to be yes: they converge around an account where individuals' intuitive systems produce multiple competing responses, and the dominant response is output with an attached evaluation of uncertainty corresponding to the relative evidence at the time of response. Importantly, this kind of account handles recent challenges to classical dual-process ideas, such as instances where traditionally second-to-complete Type II processes interfere with traditionally first-to-complete Type I processes (i.e., fast logic; Bago & De Neys, 2017; Ghasemi et al., 2022; Raoelison et al., 2021; Trippas et al., 2017). These models also mirror cue utilization models of cognition as presented earlier; meta-reasoning, as metamemory, involves the integration of multiple cues, a metacognitive evaluation of the processing of these cues, and adaptive control of one's cognition in order to optimize performance. However, metacognitive processes in the empirical work underlying these models are near universally measured in response times or confidence judgments. The logic for inferring metacognitive processes via response times is as follows: on some trials, participants took longer to respond, so they must have engaged deliberative processing, and thus they must have felt uncertain. In essence, processes are inferred by measurement of the product of their product. Alternatively, metacognition of reasoning is studied by explicit judgment; this raises questions highlighted by the theoretical and cross-cognitive work above. To wit, do explicit judgments alter the natural cognitive processes involved in decision-making as they do in other domains? This is one of the questions to which I hope to provide an answer in the present thesis.

1.4 The Present Investigation

In the present thesis, I aim to assess the degree to which explicit judgments of confidence affect metacognitive processes in reasoning. Because metacognitive processes have been so

heavily implicated in modern dual-process theories of decision-making, and because dual-process theories have been so heavily implicated in numerous consequential, real-life behaviours, this is a critical step. Throughout, I use a syllogistic reasoning task with the same set of items for each experiment. This type of task is one of the most common in the field; indeed, the majority, but not all, of the theoretical claims introduced to this point have been established in syllogistic reasoning. I use syllogisms instead of other tasks for four main reasons: i) because they are the historically dominant task, and my goal is ultimately to provide conclusions that generalize to the field, ii) they have been demonstrated to be able to differentiate between intuitive and analytic processing, the latter of which is a common operationalization of metacognitive control, iii) they have a discernibly correct answer, unlike some other common reasoning tasks, and iv) as some of our hypotheses make relatively granular claims about item characteristics, we use a set of items which has been pilot-tested with respect to those characteristics.

Evaluating the degree to which explicit judgments of confidence affect metacognitive processes in reasoning is crucial because the link to real-world behaviours implies that the mechanisms behind reasoning in standard laboratory tasks operate outside of the laboratory as well. That is, we assume in reasoning that metacognitive monitoring for internal broadcast is no different than monitoring for external report, in spite of evidence in other domains that this is not the case. Likewise, we assume that control processes are driven equally by representations of confidence irrespective of whether those representations are explicit or implicit, in spite of evidence (e.g., in memory research) that this is not the case. The current thesis thus provides a critical step towards the evaluation of metacognitive theories of reasoning and their application. There are also precursor steps to take in the reasoning domain in order to conclusively evaluate

reasoning comparably to other domains of cognition. As such, the early chapters of this thesis lay the groundwork through which the main question of explicit representation of confidence can be assessed.

In Chapter 2, I provide the to-date strongest assessment of the assumption that confidence judgments are causally related to control behaviour (specifically deliberation) in reasoning. This has been established to a reasonable degree in other domains, but evidence of causality in reasoning has proven difficult to come by. I utilize a design originally designed to test a different hypothesis, but which contained an unanticipated methodological artifact that left it uniquely able to address causality. In Experiments 1a and 1b, I demonstrate that manipulations of processing, which leave unaltered characteristics of to-be-processed stimuli, affect confidence and produce downstream effects on control behaviour that would be expected given a causal relationship between the two. In Experiments 2a and 2b, I alter this design such that the processing manipulation is directly of the control behaviour. I show, under a manipulation of control behaviour, no countereffects on confidence judgments. This pattern of effects is most consistent with a directionally causal relationship whereby confidence *causes* deliberation.

In Chapter 3, I identify a model of confidence predicted by various item and processing characteristics which, I argue, separate meaningfully distinct theoretical predictors of confidence. Across two experiments, I demonstrate that confidence is a product of i) individual cues derived from design characteristics of my chosen items, ii) behavioural variables which correspond to the totality of cues for any given item, and iii) residual, experiential characteristics of processing. This model predicts a substantial amount of variance in individuals' trial-level explicit confidence judgments. Given the conclusions of Chapter 2, the prescribed conclusion is that variables affecting confidence thereby also affect the target behaviour of deliberation. The

specification of such a model, and the separation of theoretically distinct signatures of processing allows a unique approach to addressing the main question of the present thesis, which I do in Chapter 4.

In Chapter 4, I test the hypothesis that at least some of the relationship between confidence and controlled behaviour in reasoning is driven by the requirement for participants to make an explicit judgment reflective of confidence itself. Constituent to this hypothesis is that there are monitoring differences depending on whether monitoring is undergone for the purpose of providing an explicit confidence judgment, or merely for assessing one's behaviour for the purpose of optimizing their performance. Given the conclusion in Chapter 2 that confidence is causally associated with deliberation, monitoring differences can also be assumed to impact the target behaviour, deliberation, albeit indirectly. To provide an answer to this question, I adapted a paradigm from studies of metacognition in nonhuman primates to operationalize confidence implicitly for comparison against a normal, explicit condition. I demonstrate that there are (small) differences in control behaviour by different operationalizations of confidence and (large) differences in monitoring. However, monitoring differences on the basis of explicitness of confidence seems restricted to experiential aspects of processing, rather than differences in processing item-based cues, in specific or in general.

1.4.1 Transparency and Openness

All experiments contained in the present thesis were reviewed and received ethics clearance from a University of Waterloo Research Ethics Committee. All exclusions and transformations are reported for every contained analysis, and analyses that were preregistered are labelled as such. Data and analysis scripts, as well as links to the preregistrations for Experiments 1b through 6 can be found at https://osf.io/yvmj7. Data, code, and preregistration

links are amalgamated in one location, rather than links being provided to the existent Open Science Framework pages for these experiments in case, at some future point, there are additional analyses, data, or experiments associated with these projects. This way, the provided Open Science Framework link will forever be applicable to the present thesis.

Chapter 2: Does (Under) Confidence *Cause* Deliberation?

2.1 Introduction

To this point, I have established that monitoring and control processes in reasoning approximately mirror similar processes in other areas of cognition. However, evidence in favour of one critical assumption of this work is conspicuously absent. Namely, it is assumed in several models and accounts of reasoning that deliberation arises specifically *because* of confidence. That is, the relationship between confidence and deliberation is thought to be causal in nature, and the major models in the field depend on this causality, at least implicitly. If, to the point of this thesis, *monitoring* processes are impacted by requirements for explicit confidence judgments, drawing conclusions about controlled behaviour requires one to be certain about downstream effects of monitoring processes on controlled behaviour. To this end, I will begin by attempting to establish strong evidence that such behaviour is causally driven by metacognitive monitoring, a position that is already mainstream. For example, Thompson and colleagues (2011) claim "…the initial Type 1 output is accompanied by a metacognitive judgment called the FOR, which, in turn, determines the extent of analytic engagement (p. 134)"; Pennycook and colleagues (2015) claim "If a conflict is successfully detected, however, more substantive Type 2 reasoning will be engaged. Thus, conflict detection is a bottom-up source of analytic engagement (p. 40)"; De Neys (2023), "If there is one intuition that clearly dominates, the strength difference will be high and no further deliberation will be engaged. In case the differences are more diffuse, deliberation will likewise be triggered (p. 9-10)."

Despite near universally correlational evidence in support of this claim, it is clear that the field-wide consensus is that the relationship between confidence (broadly construed) and deliberative processing is a causal one. I am not making the claim that this supposition is

unreasonable; contrarily, this supposition is reasonably made on three grounds: i) there is a large and reliable correlation between common confidence judgments and measures of deliberation, ii) it is (perhaps by far) the most parsimonious explanation for this observed relationship, and iii) metacognitive control claims in other domains are made on empirical evidence that confidence and control are causally related. Without reasonably strong evidence for causality in the reasoning and decision-making context, however, it is difficult to advance theories of meta-reasoning, especially as it pertains to the effects of monitoring processes on behaviour.

If, for example, confidence and controlled behaviour both arise from processing of a subset of the same stimulus characteristics (e.g., Kiani & Shadlen, 2009), empirical investigations of meta-reasoning may mimic causality. Empirical efforts to improve decisions via metacognition may fail, as they may rely on an absent mechanistic link between monitoring and controlled processes. Worse, they may succeed for hidden reasons, while researchers interpret them as evidence of interventions on metacognition. Research has been undertaken to address this possibility outside of a reasoning context, where similar pitfalls lie. In perhaps the best example, Desender and colleagues (2018) provide evidence that control behaviour in a perceptual decision-making task is driven by subjective confidence *and not* by performance. They had participants perform a colour discrimination task of a number of coloured shapes and manipulated the variability within the array of shapes on each trial. High variability reduced confidence, but critically did not worsen performance. Nevertheless, participants chose to review stimuli more in this condition. The argument is that this information seeking (itself a controlled behaviour) is causally driven by subjective confidence. The argued theoretical approach necessary to make this conclusion is to establish a manipulation that affects confidence *but not*

performance. It is argued that manipulations which spare performance must also have no effect on object-level representations.

Is such an approach feasible in reasoning? Almost certainly yes, but an interesting question remains as to whether the equation of performance is sufficient to claim equality of evidence in favour of one response over another. Numerous investigations of base-rate task performance (e.g., De Neys & Glumicic, 2008; Franssens & De Neys, 2009; Pennycook et al., 2012) have demonstrated participants' sensitivity to conflict between heuristic and statistical information even on trials where they ultimately respond in line with the heuristic information. This suggests there are (at least) two paths to a heuristic response: one path where individuals' monitoring system fails to detect any conflict, and one where it succeeds but analytic processing fails to engage anyway. Further yet, individuals will normatively give a heuristic response when statistical information cues the same response. It is clear from this example that task-level and metacognitive processes are not equally sensitive to stimulus characteristics; despite different object-level representations, the same response ensues. It may therefore be the case that *task performance* is not the correct outcome variable for which to control, experimentally or statistically, in order to separate metacognition from the processing of stimulus characteristics, at least in reasoning. Further to this logic, it is probably inappropriate to alter item characteristics in order to evaluate effects on monitoring and, downstream, controlled processes. After all, altered characteristics may or may not be equally represented in first-order or second-order judgments, and may be equally likely to directly impact monitoring and control behaviours.

How, then, might one manipulate confidence judgments, for the purpose of assessing downstream effects on behaviour, without controlling for performance? Perhaps a clue lies in the standard time course of reasoning. De Neys's (2023) account of the accumulation of evidence in

favour of two alternative responses suggests that participants, in an ongoing fashion, accrue evidence for each possible response. By this logic, forced continuation of processing should prolong this process and thus alter the relative evidentiary value of each response. Importantly, it would do this without an alteration to any features of items or stimuli themselves. Such a possibility is reminiscent of accounts wherein confidence judgments are informed by the continuing accumulation of evidence post-response (e.g., Van Zandt & Maldonado-Molina, 2004; Yu et al., 2015). These accounts suggest that an enforced delay might, while leaving stimulus characteristics unaltered, alter the relative evidentiary value between two alternative choices. In Experiments 1a and 1b, I use this logic to manipulate subjective confidence judgments in order to test whether these effects produce downstream effects on controlled behaviour that would be predicted by a causal association between the two.

2.2 Experiment 1a

In Experiment 1a, I test one of the component hypotheses within the hypothesis that deliberation in reasoning can be causally attributed to metacognition (more specifically, confidence). Specifically, I test the extent to which a manipulation of confidence produces the expected downstream effects on deliberation (in the opposite direction, as they are negatively correlated). Experiment 1a was originally designed to test a hypothesis similar to that in Stewart and colleagues (2023), namely that an enforced processing delay might act as a cue for difficulty (and therefore *reduce* confidence judgments). It became immediately clear that an artifact in the design rather likely prompted continued accumulation of evidence in favour of one's chosen response, as evidenced by a notable *increase* in confidence. While this left Experiment 1a unable to address the intended question, it rather conveniently rendered it appropriate to address the causal relationship between monitoring and control behaviours in reasoning. Because

Experiment 1 was not originally designed for this purpose and is not preregistered, unlike all other experiments contained in this thesis, I exactly replicate Experiment 1a in Experiment 1b.

2.2.1 Method

Participants. Thirty-eight undergraduate students from the University of Waterloo (77.4% female self-identified, M_{age} = 19.7) participated in exchange for partial course credit. This sample size was not preregistered, but a *post-hoc* sensitivity analysis suggests it provides the power to reliably (80% of the time) detect effect sizes of Cohen's d = 0.467 using a paired-samples t-test.

Design. There was one key independent variable in this experiment that had two levels (Delay vs. No Delay). Each condition had its own block, the order of which was counterbalanced across participants such that half of the participants saw each block before the other. All participants completed both conditions (i.e., the manipulation was fully within-subjects). Also counterbalanced across participants was the order of the response options; for half of participants, the yes response option appeared on the left half of the screen and the no response option on the right side, and for the other half of participants this ordering was reversed. The key dependent variables were FOR (Feeling of Rightness, an intuitive response assessing participants' perceptions of the likelihood they responded correctly; Appendix B), Rethinking Time (time to respond to the second iteration of each problem), and Answer Change (a binary variable indicating whether an individual's response changed from their first response to their second). The latter two variables are designed to capture deliberation.

Stimuli and Apparatus. Stimulus presentation and response were controlled by MATLAB 2014b using Psychophysics Toolbox 3.0.12 (Brainard, 1997; Kleiner et al., 2007). The stimuli

consisted of 48 individual syllogisms, which varied systematically in terms of their type (categorical vs. conditional), validity, and believability of their premises and conclusions (Solcz, 2011). These stimuli have been pilot-tested regarding the extent to which their premises and conclusions are believable; each item has an associated believability rating for each of its premises and its conclusion (three in total). Both of the premises of an item are always both believable or both unbelievable. Appendix A displays the items and their associated piloted believability values.

Procedure. Participants were recruited through the University of Waterloo participant pool and completed the experiment in-person in a designated testing room. Participants began by reading the consent and information form and indicating their desire to participate. Before beginning the to-be-analyzed trials, participants completed three practice syllogisms. Participants provided two responses and a confidence (FOR) judgment on each trial. Each response to each syllogism required them to choose "yes" if they determined that the syllogism was logically valid, and "no" if they determined that the syllogism was not logically valid. On the first response they gave, it was emphasized that participants were to answer as quickly as possible, and on the second response, they were instructed to take their time to ensure they gave the correct response. Each response to each syllogism was given using the mouse to click their desired response option (i.e., yes or no). The response options were equidistant (~450 pixels) from the centre of the screen, and the cursor reset to the vertical and horizontal centre of the screen before each response. On No Delay trials, the mouse operated as normal; on Delay trials, it was slowed to a speed of three pixels per frame (instituting a delay of approximately three seconds in choosing a response). Importantly, the syllogism remained on screen while the mouse was moving, irrespective of its speed. FOR judgments were given using the 1-7 keys on the

keyboard, after which participants responded to the syllogism a second time. During the second response, the mouse always moved at normal speed. This procedure was repeated for all 48 syllogisms (one block of 24 Delay trials and one block of 24 No Delay trials, in a random order), and then participants indicated their age and biological sex to complete participation.

2.2.2 Results

All data analysis and visualization was completed in the R programming language (R Core Team, 2017) with the assistance of add-on packages (see: BayesFactor [Morey & Rouder, 2024]; ez [Lawrence, 2016]; gghalves [Tiedemann, 2022]; lsr [Navarro, 2015]; plyr [Wickham, 2011]; psych [Revelle, 2024]; tidyverse [Wickham et al., 2019]), and with data visualization guidance from van Langen (2020). All analyses were carried out with an α of .05, and *t*-tests are complemented with Bayesian *t*-tests using a standard default prior width of √2/2 (Morey & Rouder, 2011). Outliers were addressed at the individual level. That is, each individual participant's trials with an RT Z-score (for either their first or second response) of greater than 3 compared with that participant's other trials were removed from analysis, and individuals whose first or second RT was a Z-score greater than 3 from the overall participant mean were also removed from analysis. This resulted in the removal of 21 trials (1.15% of trials) which were outliers at the first response and 31 trials (1.70% of trials) which were outliers at the second response. No participants were removed from analysis, and as such all analyses are based on the full sample of 38 participants.

The main set of analyses test the effect of the delay manipulation on measures of confidence (namely FOR) and, in turn, reflection (namely Rethinking Time and Answer Change). First, there was a statistically significant effect of the delay manipulation on FORs (Figure 2), $t(37) = 3.09$, $p = .004$, $d = 0.50$, $BF_{10} = 9.64$, such that FORs were higher when there

was a delay (M = 5.24, SD = 0.90) than when there was not (M = 4.90, SD = 1.10). This is the unexpected effect mentioned above, and while I am generally agnostic to the reason that a delay like the one implemented here might inflate confidence, there are clues that narrow down the mechanism. For example, it is necessary that the syllogism remain on screen while the delay occurs for this effect to exist (Stewart et al., 2023). It is therefore very likely the case that enforced continued processing extends evaluation of the syllogism along its trajectory. In any case, an account wherein confidence causally affects deliberation, as is often presumed in the literature, predicts that an increase in FORs should produce a decrease in deliberation. Instead, I observe neither a statistically significant effect on Rethinking Time (Figure 3; M_D = 10.30s, SD_D = 6.15s, M_{ND} = 10.72s, SD_{ND} = 5.90s), $t(37) = 0.57$, $p = .572$, $d = 0.09$, $BF_{01} = 10.80$, nor Answer Change (Figure 4; M_D = 17.48%, SD_D = 11.24%, M_{ND} = 17.15%, SD_{ND} = 12.80%), $t(37) = 0.17$, $p = .869$, $d = 0.02$, $BF_{01} = 5.65$. While both effects are descriptively in the same direction of what would be predicted by a causal relationship between FOR and deliberation, the effects are too small to be detected by traditional hypothesis testing. Additionally, Bayesian analysis suggests evidentiary favour of the null hypothesis in both cases. That is, Experiment 1a calls into question the causal nature of this relationship.

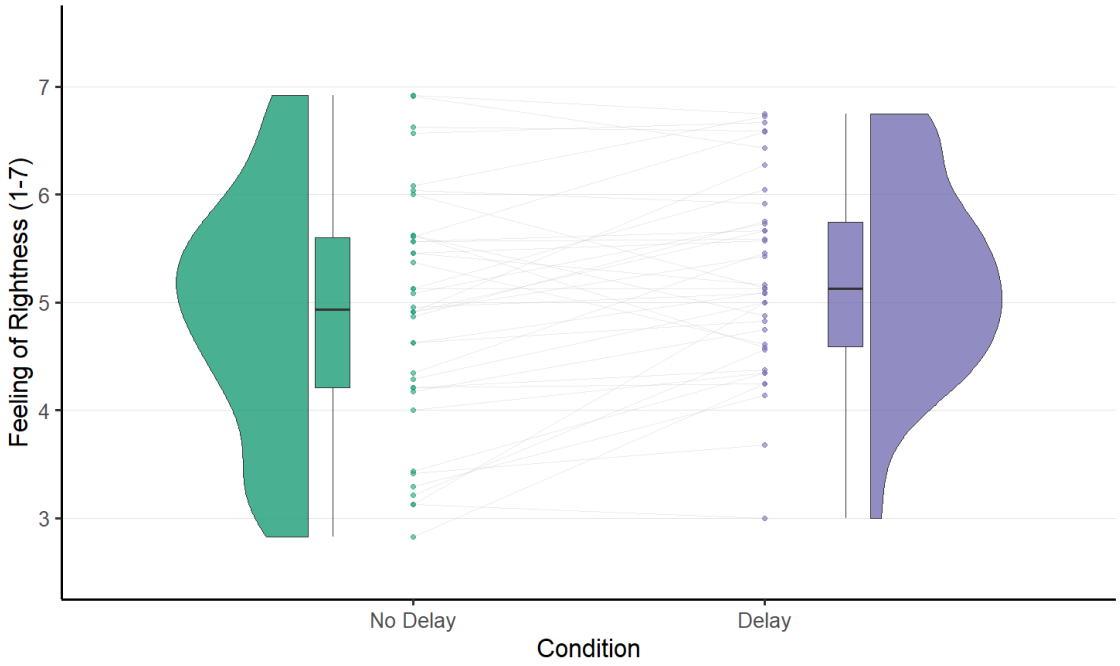


Figure 2. Feeling of Rightness (FOR) by Delay condition. Each pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles.

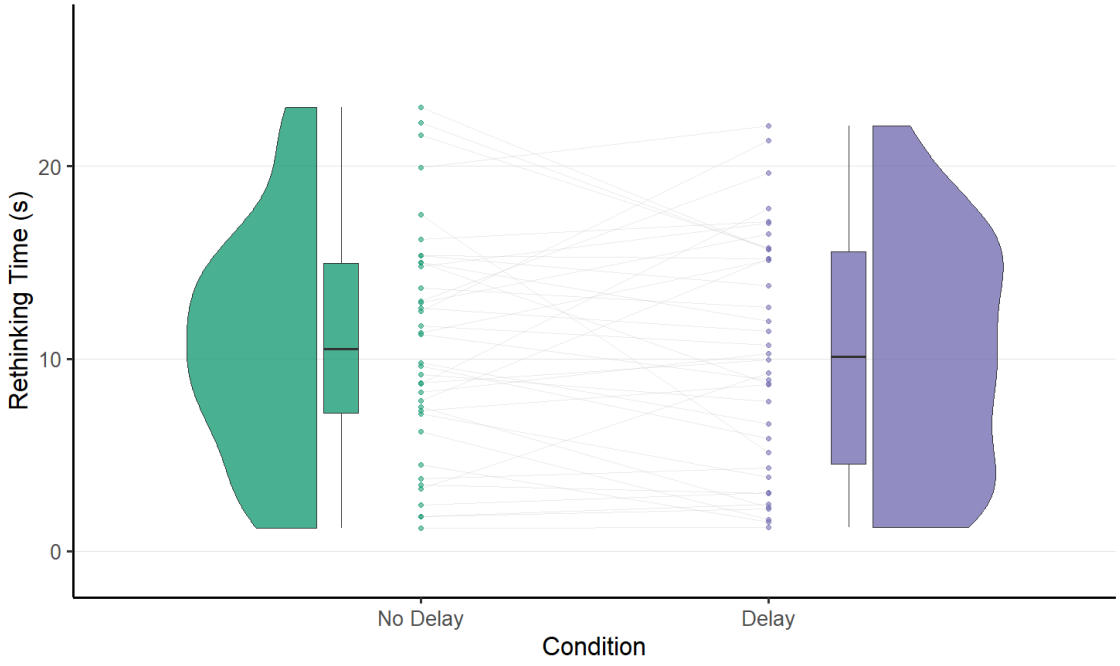


Figure 3. Rethinking Time by Delay condition. Each pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles.

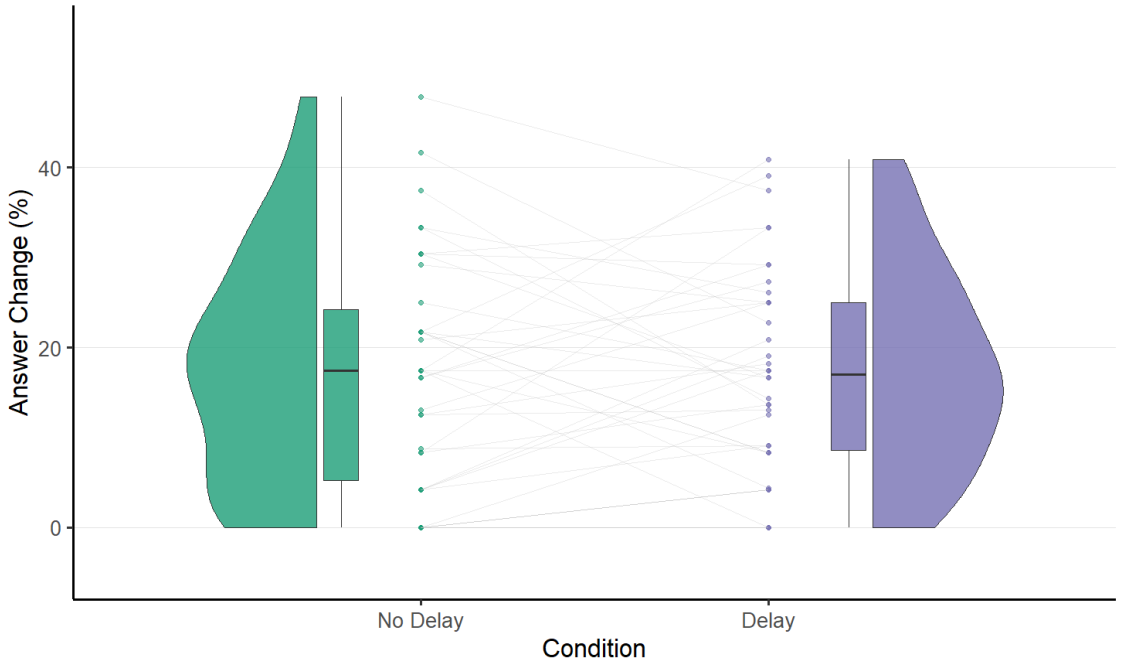


Figure 4. Answer Change by Delay condition. Each pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles.

Exploratory Analyses. Although the main analyses of Experiment 1a cast uncertainty on whether FORs and deliberation are causally related, they can only suggest a noncausal model if the two variables are still correlated (as they usually are in the literature). Otherwise, perhaps these results are due to random chance. To test this, within-subjects correlations were computed for each participant between FOR and Rethinking Time, which was significantly different from zero (r = -.40), $t(37) = 11.84$, $p < .001$, $d = 1.92$, $BF_{10} = 1.89 \times 10^{11}$. That is, it is not the case that by random chance FOR and Rethinking Time are uncorrelated, but simply that that correlation does not lend itself to a causal account of the relationship between them. Similarly, FORs were higher for unchanged answers ($M = 5.27$, $SD = 0.87$) than for changed answers ($M = 3.86$, $SD = 1.01$), $t(35) = 13.00$, $p < .001$, $d = 2.17$, $BF_{10} = 1.14 \times 10^{12}$. Note that two participants never changed an answer from Time 1 to Time 2 and are thus excluded from this analysis. These results suggest a normal pattern of relationships between FOR and deliberation, but another common

observation in reasoning is that FORs themselves are based on RT at time 1. This is true here as well, as the average within-subjects correlation between RT and FOR (r = -.14) was significantly different from zero, $t(37) = 3.73$, $p < .001$, $d = 0.61$, $BF_{10} = 46.35$. In sum, though a manipulation which affects FORs does not produce the predicted effects on deliberation, casting doubt on whether they are causally related, the data otherwise match what would be anticipated.

Another way one can test how representative these data are is to examine factors that affect FORs, as there are predictable patterns with respect to how item characteristics (e.g., syllogism believability and validity) affect metacognitive judgments like confidence. FORs were higher when syllogisms were valid ($M = 5.21$, $SD = 0.99$) as opposed to invalid ($M = 4.93$, $SD = 0.98$), $t(37) = 3.10$, $p = .004$, $d = 0.50$, $BF_{10} = 9.75$. They were higher for syllogisms that were conditional ($M = 5.25$, $SD = 0.97$) than syllogisms that were categorical ($M = 4.89$, $SD = 1.05$), $t(37) = 3.13$, $p = .003$, $d = 0.51$, $BF_{10} = 10.59$. A Premise Believability x Conclusion Believability ANOVA revealed that both Premise Believability, $F(1, 37) = 21.08$, $p < .001$, $\eta^2_g = .025$, $BF_{10} = 386.69$ and Conclusion Believability, $F(1, 37) = 5.02$, $p = .031$, $\eta^2_g = .008$, $BF_{10} = 1.2$ impacted FORs such that believable premises and conclusions were associated with higher FORs. The Premise x Conclusion Believability interaction was not statistically significant, $F(1, 37) = 2.06$, $p = .160$ $\eta^2_g = .003$, $BF_{01} = 2.25$.

2.2.3 Discussion

The results of Experiment 1a cast doubt on the notion that the relationship between confidence (in this case, FOR) and deliberation is a causal one because a manipulation that affects confidence does not produce downstream effects on deliberation. In that case, the mechanistic process that produces this relationship is unclear. It is plausibly the case that *both* FORs and deliberations are behaviours that arise from the same order of computation within the

task space. That is, FORs are produced by the processing of several aspects of the items themselves, as depicted in the last set of exploratory analyses. Could it be the case that deliberation arises in a similar fashion from the same or a similar set of predictors? That is, could the cognitive system derive *both* a signature of the likelihood that one has answered correctly *and* an indicator that more or no more processing is needed from the same set of representations (or, at least, representations all at the same level or processing)? This seems reasonable, but it is a large claim to make on the basis of a small, non-preregistered study whose aim was originally to test a different hypothesis. For this reason, Experiment 1b is a larger, preregistered, exact replication of Experiment 1a.

2.3 Experiment 1b

In Experiment 1b, I directly replicate Experiment 1a, but with a much larger sample and with preregistered hypotheses and analysis plan. Once again, I am testing the extent to which factors that impact confidence produce the expected downstream effects on deliberation. In Experiment 1a, a manipulation that affected confidence did not produce any detectable effects on deliberation, casting doubt on this claim. This would be problematic for large portions of the meta-reasoning literature, as causality is often assumed, if not directly stated.

2.3.1 Method

Participants. Eighty-five undergraduate students from the University of Waterloo (76.2% female self-identified, M_{age} = 20.2) participated in exchange for partial course credit. A sample-size of 84 was preregistered and yields approximately 95% power to detect an effect size of d = .40.

Design. Experiment 1b was an exact replication of Experiment 1a. Like Experiment 1a, there was one key independent variable that had two levels (Delay vs. No Delay). Each condition had its own block, the order of which was counterbalanced across participants such that half of participants saw each block before the other. All participants completed both conditions (i.e., the manipulation was fully within-subjects). Also counterbalanced across participants was the order of the response options; for half of participants, the yes response option appeared on the left half of the screen and the no response option on the right side, and for the other half of participants this ordering was reversed. The key dependent variables were FOR (a measure of confidence that the participant has correctly responded, measured on a 1-7 scale[1]), Rethinking Time (time to respond to the second iteration of each problem), and Answer Change (a binary variable indicating whether an individual's response changed from their first response to their second).

Stimuli and Apparatus. Stimulus presentation and response was controlled by Octave 4.4.1 using Psychophysics Toolbox 3.0.12 (Brainard, 1997; Kleiner et al., 2007). The stimuli were the same as in Experiment 1a (Appendix A).

Procedure. Participants were recruited through the University of Waterloo participant pool and completed the experiment in-person. Participants began by reading the consent and information form and indicating their desire to participate. Before beginning the to-be-analyzed trials, participants completed the same three practice syllogisms as in Experiment 1a. Participants provided two responses and a confidence (FOR) judgment on each trial. Each response to each syllogism required them to choose "yes" if they determined that the syllogism was logically valid, and "no" if they determined that the syllogism was not logically valid. On the first

[1] The preregistration states that FOR will be measured on a 100-point scale. At some point before data collection, I decided it was best that Experiment 1b be an exact replication of Experiment 1a. In error, the preregistration was not updated to reflect this. The rest of the preregistration, however, applies as normal.

response they gave, it was emphasized that participants answer as quickly as possible, and on the second response, they were instructed to take their time to ensure they gave the correct response. Each response to each syllogism was given using the mouse to click their desired response option (i.e., yes or no). The response options were equidistant (~450 pixels) from the centre of the screen, and the cursor reset to the vertical and horizontal centre of the screen before each response. On No Delay trials, the mouse operated as normal; on Delay trials, it was slowed to a speed of three pixels per frame (instituting a delay of approximately three seconds in choosing a response). Importantly, the syllogism remained on screen while the mouse was moving, irrespective of its speed. FOR judgments were given on a seven-point scale by button press (Appendix B), after which participants responded to the syllogism a second time. During the second response, the mouse always moved at normal speed. This procedure was repeated for all 48 syllogisms (one block of 24 Delay trials and one block of 24 No Delay trials), and then participants indicated their age and biological sex to complete participation.

2.3.2 Results

All data analysis and visualization was completed in the R programming language (R Core Team, 2017) with the assistance of add-on packages (see: : BayesFactor [Morey & Rouder, 2024]; ez [Lawrence, 2016]; gghalves [Tiedemann, 2022]; lsr [Navarro, 2015]; plyr [Wickham, 2011]; psych [Revelle, 2024]; tidyverse [Wickham et al., 2019]), and with data visualization guidance from van Langen (2020). All analyses were carried out with an α of .05, and *t*-tests are complemented with Bayesian *t*-tests using a prior width of $\omega = .2$. Eighty percent of this prior distribution corresponds to effect sizes between -.59 and .59 (see Schmalz et al., 2023), which is the upper bound of the 60% confidence interval of the Cohen's d effect observed in the to-be-replicated experiment. Outliers were addressed at the individual and at the overall level. That is,

each individual participant's trials with an RT Z-score (for either their first or second response) of greater than 3 compared with that participant's other trials were removed from analysis, and individuals whose mean value on any other variable of interest was a Z-score greater than 3 from the overall participant mean were also removed from analysis. This resulted in the removal of 43 trials (1.03% of trials) which were outliers at the first response and 64 trials (1.53% of trials) which were outliers at the second response. One participant was removed from analysis for being an outlier with respect to Rethinking Time, leaving 84 participants. All following analyses are based on these 84 participants.

Preregistered Analyses. I first analyze the effect of the delay manipulation on measures of FORs and, in turn, reflection (as approximated by Rethinking Time and Answer Change). FORs were greater when there was a delay ($M = 4.79$, $SD = 1.01$) then when there was not ($M = 4.62$, $SD = 0.98$), $t(83) = 2.45$, $p = .017$, $d = 0.27$, $BF_{10} = 3.32$ (Figure 5). This is a replication of the same effect in Experiment 1a, although noticeably smaller. However, the crucial question is whether once again this effect is accompanied by the absence of an offsetting effect on measures of deliberation. This time, there is a statistically significant effect on Rethinking Time (Figure 6), $t(83) = 2.14$, $p = .035$, $d = 0.23$, $BF_{10} = 1.96$, such that Rethinking Time was lower when there was a delay ($M = 13.25$s, $SD = 7.10$s than when there was no delay ($M = 14.66$, $SD = 7.91$s). There was not, however, a statistically significant effect on Answer Change (Figure 7; $M_D = 17.48\%$, $SD_D = 12.49\%$, $M_{ND} = 20.37\%$, $SD_{ND} = 11.20\%$), $t(83) = 1.94$, $p = .056$, $d = 0.21$, $BF_{10} = 1.44$. While both effects are in the predicted direction and approximately the same size (and approximately equal to the effect on FORs), only one reaches statistical significance. Ultimately, it feels a reasonable conclusion to suggest that in Experiment 1b, unlike Experiment 1a, the effect of the delay manipulation on FORs is matched by effects on deliberation. That is, the two

experiments are qualitatively dissimilar, and the preregistered one with a larger sample is generally consistent with the view that FORs *cause* deliberation.

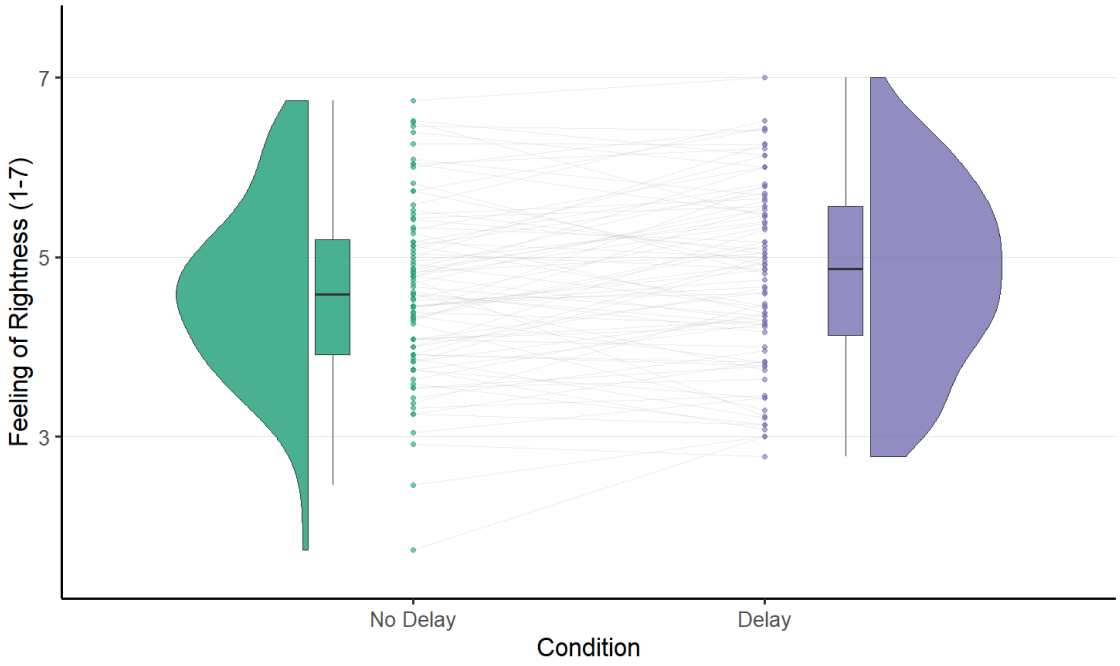


Figure 5. Feeling of Rightness (FOR) by Delay condition. Each pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles.

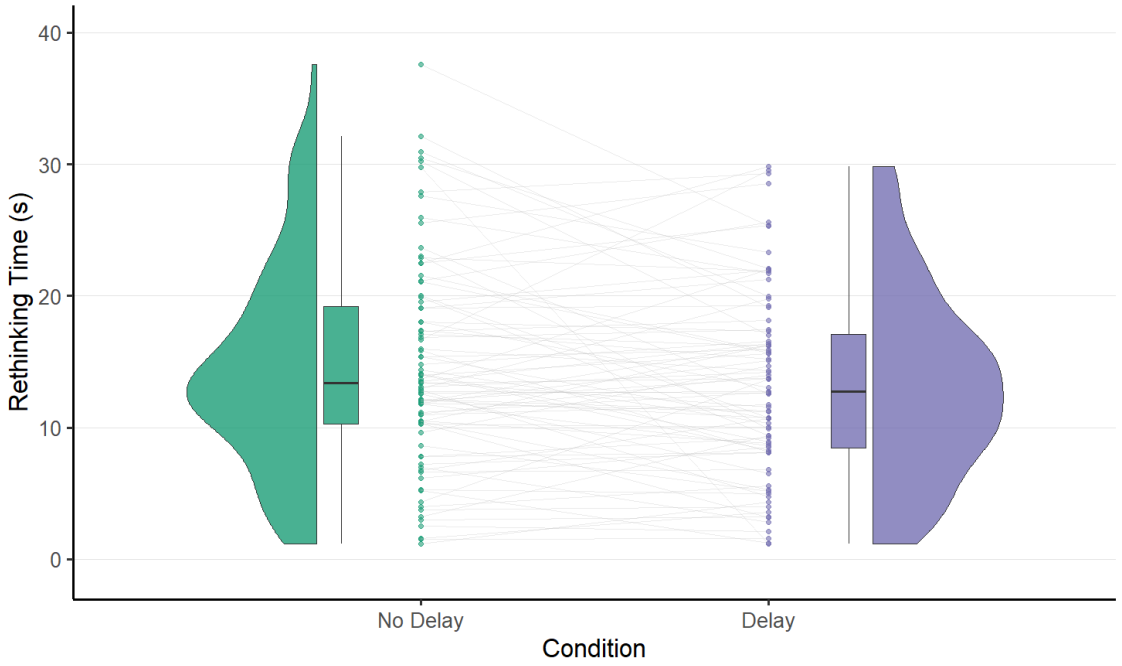


Figure 6. Rethinking Time by Delay condition. Each pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles.

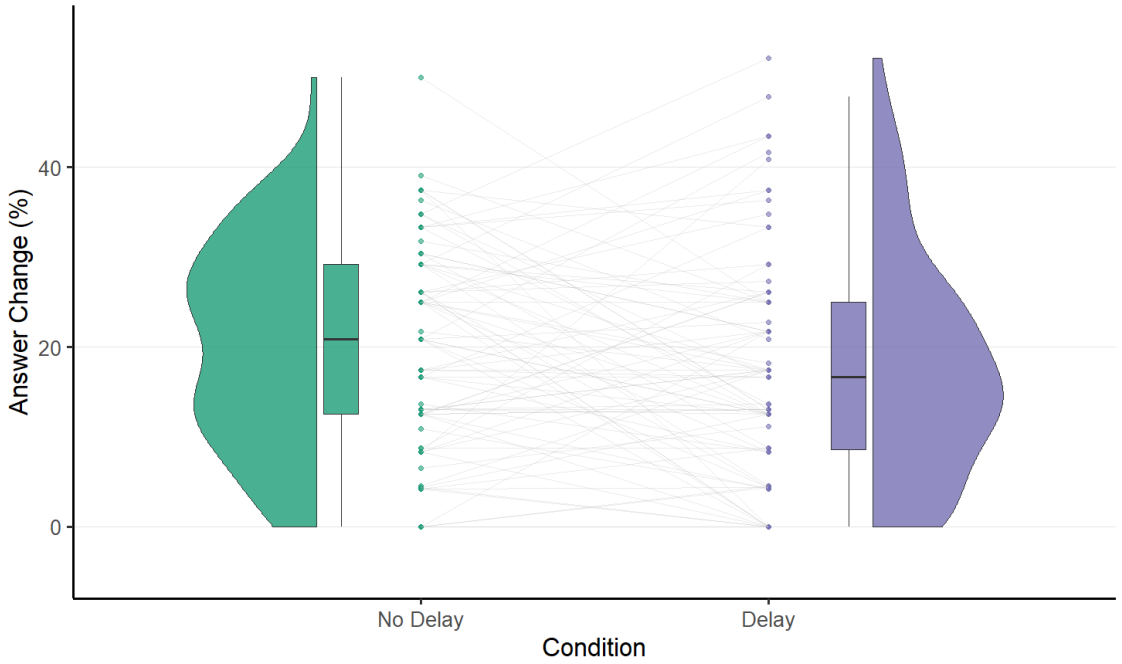


Figure 7. Answer Change by Delay condition. Each pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles.

Exploratory Analyses. Exploratory analyses were conducted to mirror those conducted in Experiment 1a. The average within-subject correlation between FOR and Rethinking Time (r = -.42) was statistically significant, $t(83) = 18.18$, $p < .001$, $d = 1.98$, $BF_{10} = 8.83 \times 10^{26}$. FORs for trials on which participants changed their answers ($M = 3.39$, $SD = 1.08$) were significantly lower than FORs for trials where participants did not change their answers ($M = 4.99$, $SD = 0.91$), $t(82) = 18.03$, $p < .001$, $d = 1.98$, $BF_{10} = 3.42 \times 10^{26}$. One participant never changed their answer over the course of the experiment and has been excluded from this analysis. The average within-subjects correlation between RT and FOR was significant (r = -.15), $t(83) = 6.16$, $p < .001$, $d = 0.67$, $BF_{10} = 285{,}943$.

As in Experiment 1a, I test the effects of various item characteristics on FORs. FORs were again higher when syllogisms were valid ($M = 4.92$, $SD = 0.95$) as opposed to invalid ($M = 4.49$, $SD = 0.99$), $t(83) = 8.07$, $p < .001$, $d = 0.86$, $BF_{10} = 4.24 \times 10^{8}$. They were higher for syllogisms that were conditional ($M = 4.85$, $SD = 0.95$) than syllogisms that were categorical ($M = 4.57$, $SD = 1.02$), $t(83) = 4.55$, $p < .001$, $d = 0.50$, $BF_{10} = 784.12$. A Premise Believability x Conclusion Believability ANOVA revealed that both Premise Believability, $F(1, 83) = 77.66$, $\eta^2_g = .045$, $BF_{10} = 1.90 \times 10^{12}$ and Conclusion Believability, $F(1, 83) = 27.48$, $\eta^2_g = .014$, $BF_{10} = 663.23$ impacted FORs such that believable premises and conclusions were associated with higher FORs. The Premise x Conclusion Believability interaction was not statistically significant, $F(1, 83) = 1.98$, $p = .163$, $\eta^2_g = .001$, $BF_{01} = 1.27$.

Experiment 1a and Experiment 1b yield essentially opposite conclusions vis-à-vis the nature of the relationship between FORs and deliberation. In Experiment 1a, a manipulation that produces a statistically significant effect on FORs produces a much smaller effect on deliberation. In Experiment 1b, both effects are roughly equivalent (however, only one of two

candidate effects on deliberation reaches statistical significance). That said, the effects are all directionally consistent across experiments. As such, I repeat the main analyses with the combined data from both experiments to estimate these effects with the full available data. In the combined data, FORs were greater when there was a delay (M = 4.93, SD = 1.00) than when there was not (M = 4.71, SD = 1.02), $t(121) = 3.76$, $p < .001$, $d = 0.34$, $BF_{10} = 87.55$ (Figure 8). Rethinking Time was lower when there was a delay (M = 12.33s, SD = 6.93s) than when there was No Delay (M = 13.44s, SD = 7.54s), $t(121) = 2.17$, $p = .032$, $d = 0.20$, $BF_{10} = 2.00$ (Figure 9). There was no statistically significant effect on Answer Change (Figure 10; M_D = 17.48%, SD_D = 12.07%, M_{ND} = 19.37%, SD_{ND} = 11.76%), $t(121) = 1.57$, $p = .120$, $d = 0.14$, $BF_{01} = 1.22$. Because the effect of the manipulation on FORs also produces opposite effects on at least one of the measures of deliberation (and directionally similar effects on both), I take this as evidence in support of the hypothesis that the two are causally related.

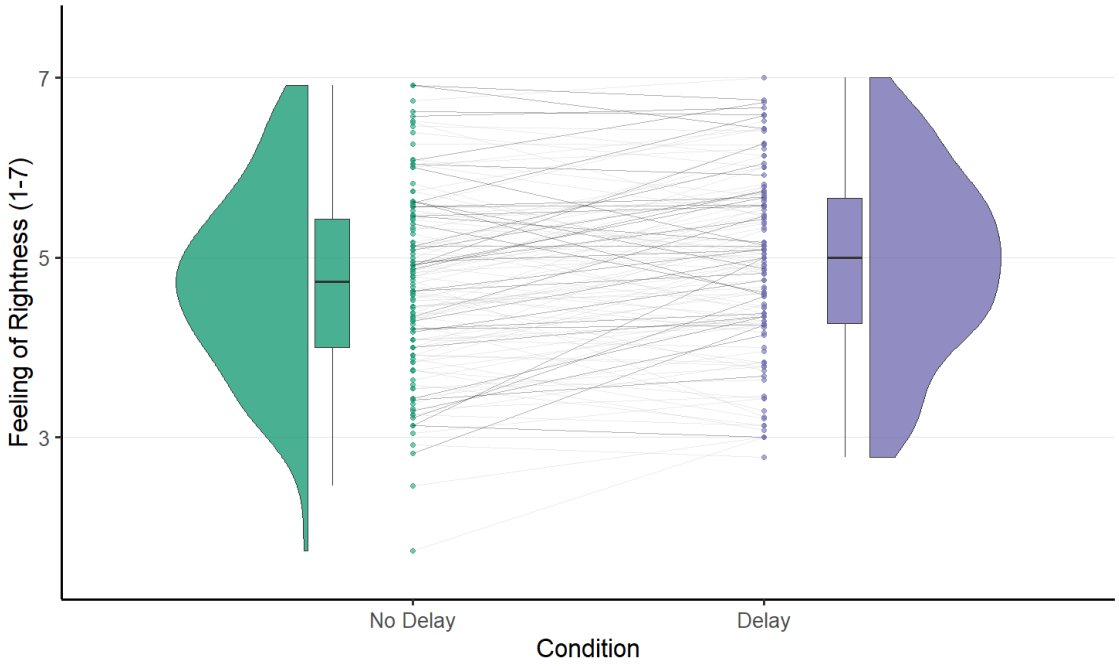


Figure 8. Feeling of Rightness (FOR) by Delay condition. Each pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles. Darker slopes indicate data from Experiment 1a.

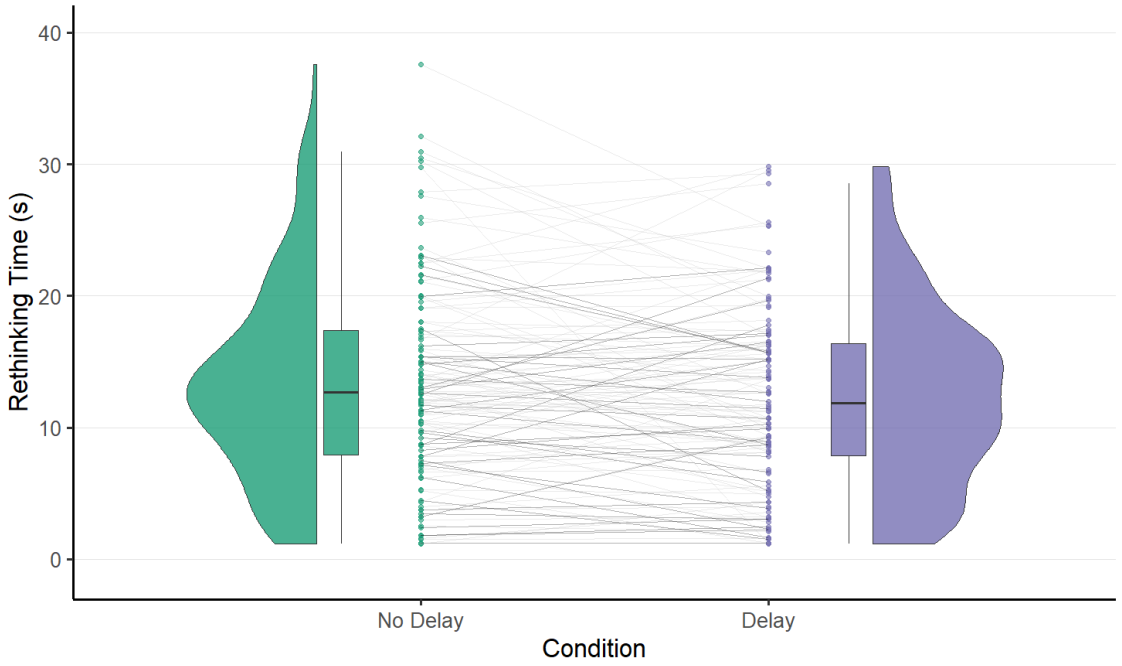


Figure 9. Rethinking Time by Delay condition. Each pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles. Darker slopes indicate data from Experiment 1a.

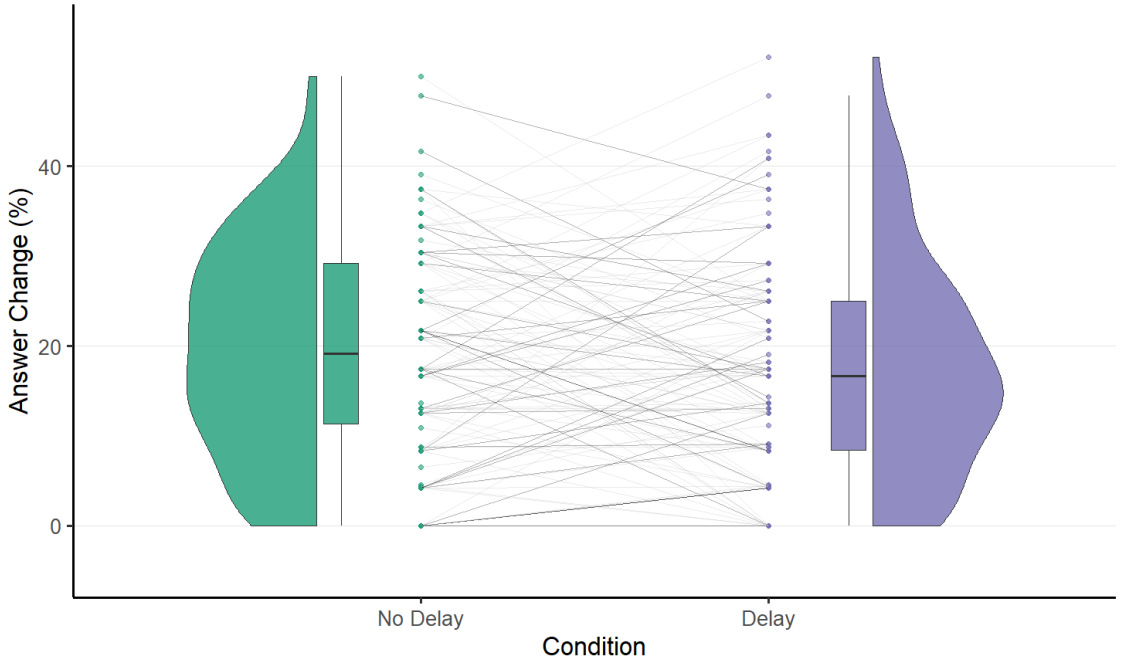


Figure 10. Answer Change by Delay condition. Each pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles. Darker slopes indicate data from Experiment 1a.

2.3.3 Discussion

In Experiment 1b, I show evidence for a causal relationship between confidence (as measured by FOR judgments) and deliberation; specifically, a manipulation that increases confidence in the aggregate decreases Rethinking Time in the aggregate. Answer Change also decreases, but to a level that fails to reach statistical significance in a standard analysis. These results are somewhat inconsistent with the results of Experiment 1a, wherein there was not a detectable effect on either measure of reflection. It is difficult to establish a reason the results from these experiments differ, but inspection of Figures 9 and 10 provides a suggestion that there was overall less reflection in Experiment 1a than in Experiment 1b. Perhaps the difference is random; perhaps the cause is invariance in the variables capturing deliberation, but in any case, it raises the question of which set of results to believe; given the larger sample and preregistered nature of Experiment 1b, it seems the easy choice. Additionally, the effects of the manipulation on confidence and both measures of deliberation are approximately the same in Experiment 1b (effect sizes ranging from $d = 0.21$ to $d = 0.27$), while they vary wildly in Experiment 1a (ranging from $d = 0.02$ to $d = 0.50$). Alternatively, one could evaluate Experiments 1a and 1b in tandem. The combined results are largely consistent with the results of Experiment 1b, but also the results of Experiment 1a themselves are directionally consistent with the results of Experiment 1b. That is, any way one slices it, the results are relatively consistent with a causal account of the relationship between FOR and deliberation. Not to mention, deliberation is multiply determined, not solely by confidence. In some ways, it is unsurprising that the effects on deliberation are smaller than those on confidence.

Unclear is the actual mechanistic effect the delay manipulation has on confidence and the reasoning process. Recall that, for the delay to affect confidence, the syllogism must remain on-

screen while the delay occurs (Stewart et al., 2023). One can think of reasoning as a process of accumulating evidence in favour of the two alternative options. I will explore an Evidence Accumulation account of the metacognition of reasoning in more depth in Chapter 3. In at least some cases, evidence strength for one's ultimate choice will be increasing relative to the ultimately unchosen choice (i.e., they will be diverging) at the time of a decision. A delay on deciding, if indeed that is what the delay is doing, may result in further divergence between alternative choices and increased confidence in the resultant judgment. Though I have no direct evidence for any mechanism, this is the kind of explanation I favour, as it is consistent with the results herein and also with modern major theories of dual-process reasoning (see De Neys, 2023). However, there are some interesting corollaries to this kind of account; for example, what if the relative evidence for two alternative choices is converging at the time of a decision? One would predict that an enforced delay would, via extending the accumulation of evidence along its trajectory, actually *decrease* confidence, a pattern that is apparent for some participants in both Experiments 1a and 1b. If there are individual differences that predict, say, the order in which cues are processed such that some individuals generally have converging patterns of evidence strength, a delay manipulation will repeatably decrease their confidence (as measured by Feelings of Rightness). These are interesting questions, but not the focus of the present thesis. In any case, the delay manipulation prompts additional pre-response accrual of evidence in favour of each possible response. While this is by some standards (e.g., the one set by Desender et al., 2018) enough to establish causality, it also should be the case that manipulations affecting the outcome variable do not impact the argued causal variable. That is, it is still theoretically possible, given the results of Experiments 1a and 1b, that the manipulation is really of task-level representations. I address this possibility in Experiments 2a and 2b.

2.4 Experiment 2a

In Experiments 1a and 1b, I tested one component hypothesis to the claim that FOR causally affects deliberation: that manipulations affecting FORs also affect deliberation, for which I find evidence (although somewhat mixed evidence). This is a necessary but not sufficient condition to establish causality; there are plausible models where the delay manipulation in Experiment 1a simply altered the processing of the feature space and this impacted *both* confidence and deliberation in the same way. That is, it is still plausible that confidence and deliberation are behaviours at the same level (i.e., based on the same inputs), rather than one being caused by the other (i.e., confidence being an input to deliberation). However, there is a relatively simple prediction that can exclude this kind of account: any manipulation that targets deliberation should not impact FORs. If one can demonstrate that deliberation is directly manipulatable and that successful manipulations of deliberation do not also affect confidence, it would rule out at least some of this family of counterclaims. There are several possible candidate manipulations one could employ, but one possibility allows for the use of a very similar manipulation to that of Experiments 1a and 1b. While the delay to the speeded response used above results in forced extended processing prior to a response, delays – or otherwise increased processing time – have also been used to directly increase the degree of analytical processing that is deployed (Evans & Curtis-Holmes, 2005; Evans, 2019). Using this logic, I target deliberation directly (independent of confidence judgments) by using a very similar methodological approach to before but with the exception of restricting the delay to the time following an initial response – but still before confidence judgments are made.

2.4.1 Method

Participants. Forty-two undergraduate students from the University of Waterloo participated in exchange for partial course credit. Demographic information for this experiment is not available. This sample size was preregistered based on a power analysis conducted in G*Power 3 (Faul et al., 2007)[2]. A simulation-based sensitivity analysis conducted in the R programming language using the *Superpower* add-on package[3] indicates that this sample size provides approximately .8 power to detect effect sizes of $\eta^2_p = .012$ or greater.

Design. There was one key independent variable that had three levels (Delay vs. Empty Delay vs. Exposure Delay). Trial types were randomly intermixed, and it was randomly counterbalanced across participants which third of items were assigned to each condition such that each third of items appeared equally often in each condition. All participants completed trials of all three types (i.e., the manipulation was fully within-subjects). Also counterbalanced across participants was the order of the response options; for half of participants, the yes response option appeared on the left half of the screen and the no response option on the right side, and for the other half of participants this ordering was reversed. The response key associated with each option was similarly counterbalanced. The key dependent variables were FOR (Feeling of Rightness, a measure of confidence that the participant has correctly responded, measured on a 7-point scale) and Answer Change (a binary variable indicating whether an individual's response changed from their first response to their second).

---

[2] I no longer do power analysis using G*Power because I believe it sometimes introduces error for studies including many trials and within-subjects manipulations (see Giner-Sorolla et al., 2024 for a broad overview). As such, rather than reporting the details of a (potentially imprecise) power analysis, I conduct a better-founded *sensitivity* analysis, which assesses the size of effect we can detect at this sample size and desired power level. This is more consistent with other experiments in this thesis.

[3] This sensitivity analysis was conducted based on 10,000 simulations using the means and standard deviations of the comparison of FOR across delay conditions in Experiment 1b, using the actual cross-condition correlation of $r = .79$.

Stimuli and Apparatus. Stimulus presentation and response was controlled by MATLAB 2014b[4] using Psychophysics Toolbox 3.0.12 (Brainard, 1997; Kleiner et al., 2007). The stimuli consisted of 48 individual syllogisms, which varied systematically in terms of their type (categorical vs. conditional), validity, and believability of their premises and conclusions (Solcz, 2011). The stimuli are again the same stimuli as in the previous experiments (Appendix A).

Procedure. Participants were recruited through the University of Waterloo participant pool and completed the experiment in-person in a designated testing room. Participants began by reading the consent and information form and indicating their desire to participate. Before beginning the to-be-analyzed trials, participants completed three practice syllogisms. This experiment utilized the same two-response paradigm as the previous experiments wherein participants responded twice to each syllogism, first speeded and then unspeeded. Between the two responses, they gave a FOR judgment. This time, participants responded to each syllogism using the keyboard, pressing the '1' or '3' key to respond in either the affirmative or negative.

On No Delay trials, the experiment proceeded immediately to the FOR judgment when participants gave their response to the first iteration of each problem. On Empty Delay trials, a three second delay (to approximately match the length of the delays in Experiments 1a and 1b) was added prior to the appearance of the FOR question, during which the syllogism was removed from the screen. On Exposure Delay trials, an equal length delay was added during which the syllogism remained on screen. Critically, the delay in all cases occurs after a participant has responded to the syllogism but before they have given a FOR judgment. This procedure was

[4] Eagle-eyed readers may notice a shift back to collecting these data in MATLAB (after switching to Octave in Experiment 1b). In the interest of full disclosure, Experiment 1b was conducted after Experiment 2a, but ordered such that it appears in this thesis next to the study it is replicating. I have moved to collecting data in Octave in favour of open science, as it is an open-source software available without a licensing fee.

repeated for all 48 syllogisms (16 syllogisms of each trial type). Because the delay occurs after the selection of a response, it allows for the possibility that the given response is accompanied by a Feeling of Rightness based on the evidence at the time of response. In this case, a delay might prompt deliberation in absence of an effect on confidence. If deliberation and Feelings of Rightness both arise from the same stimulus processing characteristics, we must expect them to be affected by the manipulation in opposite directions (or, of course, not at all). Conversely, if deliberation is a causal product of confidence, any direct manipulation of deliberation need not retroactively impact confidence.

2.4.2 Results

All data analysis and visualization was completed in the R programming language (R Core Team, 2017) with the assistance of add-on packages (see: BayesFactor [Morey & Rouder, 2024]; ez [Lawrence, 2016]; gghalves [Tiedemann, 2022]; lsr [Navarro, 2015]; plyr [Wickham, 2011]; psych [Revelle, 2024]; tidyverse [Wickham et al., 2019]), and with visualization guidance from van Langen (2020). All analyses were carried out with an α of .05, and are complemented with Bayesian analyses using a standard default prior width of $\sqrt{2}/2$ (Morey & Rouder, 2011). Outliers were addressed at the individual level. Each individual participant's trials with an RT Z-score (for either their first or second response) of greater than 3 compared with that participant's other trials were removed from analysis, and individuals whose first or second RT was a Z-score greater than 3 from the overall participant mean were also removed from analysis. This resulted in the removal of 21 trials (1.06% of trials) which were outliers at the first response and 28 trials (1.42% of trials) which were outliers at the second response. One participant was removed for withholding a response to too many trials (they only had five trials on which they provided a response). All analyses are based on the resultant 41 participants.

Preregistered Analyses. There was no statistically significant effect of the delay manipulation on FORs (Figure 11), $F(2,80) = 0.32$, $p = .730$, $\eta^2_g = .00$, $BF_{01} = 17.53$. There was, however, a statistically significant effect of the delay manipulation on Answer Change (Figure 12), $F(2,80) = 3.88$, $p = .032$, $\eta^2_g = .025$, $BF_{10} = 1.41$, but note that the accompanying Bayesian analysis suggests essentially equal evidence for a non-effect. Follow-up paired-samples *t*-tests were conducted to investigate the significant omnibus Answer Change ANOVA, and they revealed that answers were changed more often in the Exposure Delay Condition ($M = 27.41\%$, $SD = 16.05\%$) than in the No Delay condition ($M = 27.41\%$, $SD = 16.05\%$), $t(40) = 2.80$, $p = .008$, $d = 0.44$, $BF_{10} = 5.49$. Answer Change was not greater than in the Empty Delay condition ($M = 22.77\%$, $SD = 12.35\%$), $t(40) = 1.52$, $p = .137$, $d = 0.28$, $BF_{01} = 1.07$, and nor did the No Delay and Empty Delay conditions differ from one another, $t(40) = 1.19$, $p = .240$, $d = 0.19$, $BF_{01} = 1.41$. Note that the mirroring analyses with Rethinking Time were not preregistered as they are difficult to interpret. All trials in the Exposure Delay condition have three seconds more time with the syllogism after an initial response is given. If one included this time in Rethinking Time, there would be an enormous effect, but it would obviously be spurious. However, measuring Rethinking Time as normal and analyzing it across conditions might exclude some genuine deliberation that happens while the syllogism remains on-screen during the delay. As such, this analysis is not treated as a main one for the current experiment. The average within-subject correlation between FOR and Rethinking time was approximately the same as in the previous experiments ($r = -.31$), and was once again significant, $t(40) = 14.24$, $p < .001$, $d = 2.22$, $BF_{10} = 1.89 \times 10^{14}$. The average within-subjects correlation between Response Time and FORs ($r = -.16$) was also statistically significant, $t(40) = 5.29$, $p < .001$, $d = 0.83$, $BF_{10} = 4{,}058$. FORs were also

lower for answers that were changed ($M = 3.62$, $SD = 1.01$) than those that were not changed ($M = 4.62$, $SD = 0.69$), $t(40) = 8.20$, $p < .001$, $d = 1.28$, $BF_{10} = 9.66 \times 10^6$.

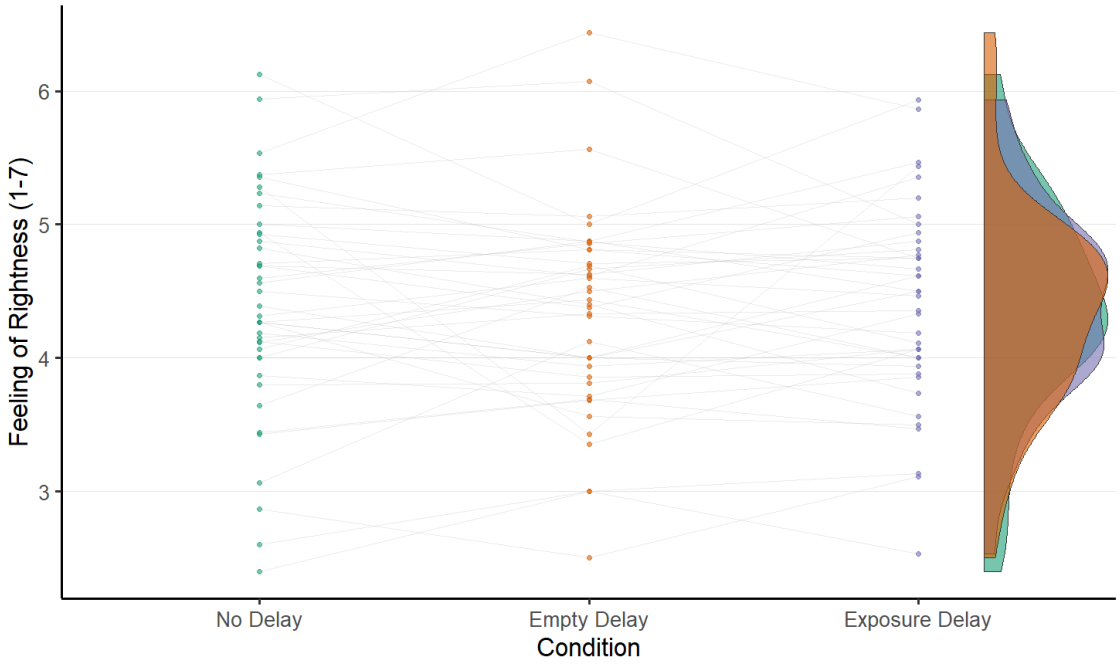


Figure 11. Feeling of Rightness (FOR) by Delay condition. Each connected set of points represents an individual participant.

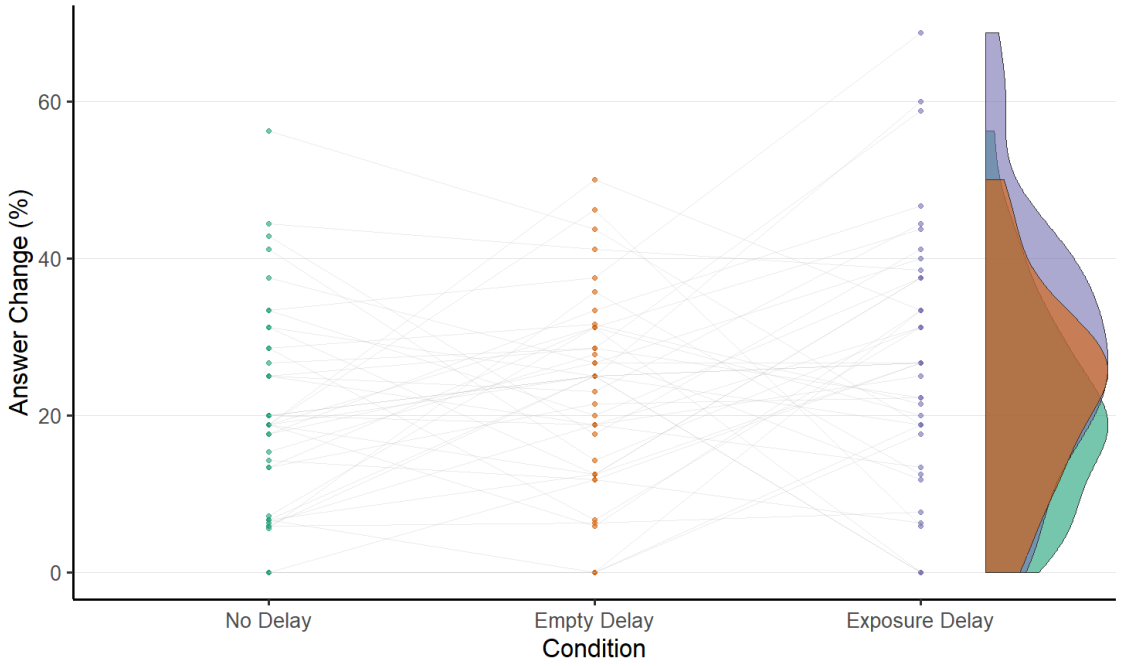


Figure 12. Answer Change (%) by Delay condition. Each connected set of points represents an individual participant.

Exploratory Analyses. First, I mirrored the above tests with respect to Rethinking Time. There was no statistically significant effect of the delay manipulation on Rethinking Time, $F(2,80) = 1.04$, $p = .358$, $\eta^2_g = .002$, $BF_{01} = 9.29$. Note that, because approximately three seconds is added to every Delay trial and there is no statistically significant effect here, any rethinking while the delay is occurring would indicate there being more deliberation overall in the Delay condition. As in Experiments 1a and 1b, the various item characteristics predictably impact FOR judgments. FORs were again higher when syllogisms were valid ($M = 4.44$, $SD = 0.76$) as opposed to invalid ($M = 4.31$, $SD = 0.72$), $t(40) = 2.21$, $p = .033$, $d = 0.35$, $BF_{10} = 2.14$. They were higher for syllogisms that were conditional ($M = 4.53$, $SD = 0.74$) than syllogisms that were categorical ($M = 4.22$, $SD = 0.77$), $t(40) = 4.08$, $p < .001$, $d = 0.64$, $BF_{10} = 84.20$. A Premise Believability x Conclusion Believability ANOVA revealed that both Premise Believability, $F(1, 40) = 31.08$, $p < .001$, $\eta^2_g = .069$, $BF_{10} = 44{,}929$ and Conclusion Believability, $F(1, 40) = 13.31$, $p < .001$, $\eta^2_g = .031$, $BF_{10} = 29.08$ impacted FORs such that believable premises and conclusions

were associated with higher FORs. The Premise x Conclusion Believability interaction was not statistically significant, $F(1, 40) = 1.15$, $p = .289$, $\eta^2_g = .002$, $BF_{01} = 1.63$.

2.4.3 Discussion

In Experiment 2a, I demonstrate, with a very similar manipulation to that in Experiments 1a and 1b, that it is possible to manipulate reflection without impacting confidence. In this case, different than the previous experiments, reflection is operationalized only as answer change, which is both necessary and problematic. It is necessary because the manipulation makes it impossible to know when one starts deliberating on a previous choice. However, there is a suggestion in the data that it begins while the syllogism remains on-screen during the enforced delay, namely that the Empty Delay does not differ vis-à-vis any dependent variable from the No Delay condition. That is, it is the presence of the syllogism onscreen that prompts deliberation, and so one can argue that means deliberation begins during the delay. If that is true, it is certainly the case that more rethinking takes place in the Exposure Delay condition than in the other two conditions, despite the absence of a hypothesis test to this effect.

Most critically, it is possible to increase deliberation without an opposite effect on confidence. This is a critical portion to the argument that confidence causes reflection. If current accounts are based on, for example, a reality where both confidence and deliberation arise from the same stimulus space, the results of Experiment 2a are very unlikely. I will discuss how it could still be possible in the Chapter 2 Discussion, but to this point, I take the current experiment as evidence to pre-empt this claim. However, Experiment 2a provides, at times, somewhat weak statistical evidence for this claim. Although it is preregistered unlike Experiment 1a, it is based on a somewhat small sample size and as a result, rests on middling evidence. As a result, I replicate Experiment 2a in a larger sample in Experiment 2b.

2.5 Experiment 2b

In Experiment 2a, I provide evidence that manipulations of deliberation do not necessarily also impact confidence. I argue this is helpful to make a causal claim about the relationship between these two variables, as it is even less likely in light of these data that this relationship is due to each arising from characteristics of processing the stimuli themselves. However, like Experiment 1a, the evidence for this claim is somewhat weak; effects hover around significance and accompanying Bayesian analyses indicate no strong evidence in either direction. Additionally, there are three levels to the independent variable and while this perhaps allows greater insight into the potential mechanism at play, it is neither necessary nor helpful to answer the specific question being asked. Regardless of whether it is necessary for stimuli to remain on screen during the delay, the critical point is that the delay impacts deliberation. As a result, in Experiment 2b I remove the disjunction between Empty and Exposure Delays. For these reasons, Experiment 2b was conducted with a larger sample size and simplified design to hopefully increase the strength of evidence that manipulations affecting deliberation do not affect confidence.

2.5.1 Method

Participants. One hundred and four undergraduate students from the University of Waterloo participated in exchange for partial course credit (75.0% female self-identified, M_{age} = 19.4). A sample size of 95 was preregistered and provides .80 power to detect effect sizes up to d = 0.29 with a paired-samples t-test. Participants are dropped as outliers or excluded at random to get to this number, and all analyses herein are based on the resultant 95 participants.

Design. The same independent variable as in Experiment 2a was included here (i.e., Delay); however, only the No Delay and Exposure Delay conditions were included. Since these

are the two conditions with the largest effect between them, and the middle condition is not necessary to make any claim about the effects of a delay on reflection nor on confidence, I exclude it in Experiment 2b. Additionally, this allows for more trials per condition (24 rather than 16), reducing the potential for noise in the condition-level group means. Trial types were again randomly intermixed, and it was randomly counterbalanced across participants which half of items were assigned to each condition such that each half of items appeared equally often in each. This manipulation was again completely within-subjects. Also counterbalanced across participants was the order of the response options; for half of participants, yes appeared on the left and was mapped to the '1' key, with 'no' on the right mapped to the '3' key, and for the other half this mapping was reversed. The key dependent variables were FOR (Feeling of Rightness, a measure of confidence that the participant has correctly responded, measured on a 7-point scale) and Answer Change (a binary variable indicating whether an individual's response changed from their first response to their second).

Stimuli and Apparatus. Stimulus presentation and response was controlled by Octave 4.4.1 using Psychophysics Toolbox 3.0.12 (Brainard, 1997; Kleiner et al., 2007). The stimuli consisted of 48 individual syllogisms, which varied systematically in terms of their type (categorical vs. conditional), validity, and believability of their premises and conclusions (Solcz, 2011). The stimuli are again the same stimuli as in the previous experiments (Appendix A).

Procedure. Participants were recruited through the University of Waterloo participant pool and completed the experiment in-person in a designated testing room. Participants began by reading the consent and information form and indicating their desire to participate. Before beginning the to-be-analyzed trials, participants completed three practice syllogisms. This experiment utilized the same two-response paradigm as the previous experiments wherein

participants responded twice to each syllogism, first speeded and then unspeeded. Between the two responses, they gave a FOR judgment. Participants responded using the keyboard, pressing the '1' or '3' key to respond in either the affirmative or negative. On No Delay trials, the experiment proceeded immediately to the FOR judgment when participants gave their response to the first iteration of each problem. On Delay trials, a three second delay was added during which the syllogism remained on screen. Critically, the delay in all cases occurs after a participant has responded to the syllogism but before they have given a FOR judgment. This procedure was repeated for all 48 syllogisms (24 syllogisms of each trial type). Recall that in an account where deliberation is caused in part by confidence, a delay manipulation directly targeting deliberation need not produce a commensurate effect on confidence.

2.5.2 Results

All data analysis and visualization was completed in the R programming language (R Core Team, 2017) with the assistance of add-on packages (see: BayesFactor [Morey & Rouder, 2024]; ez [Lawrence, 2016]; gghalves [Tiedemann, 2022]; lsr [Navarro, 2015]; plyr [Wickham, 2011]; psych [Revelle, 2024]; tidyverse [Wickham et al., 2019]), and with data visualization guidance from van Langen (2020). All analyses were carried out with an α of .05 and are complemented with Bayesian analyses using a custom prior width parameter of $\omega = .2$ which corresponds to a distribution, 80% of which encapsulates effect sizes of $d = -.57$ to $d = .57$, which is the upper bound of the 60% confidence interval of the Cohen's d effect observed in the to-be-replicated experiment. Outliers were addressed at the individual and group level. Each individual participant's trials with an RT Z-score (for either their first or second response) of greater than 3 compared with that participant's other trials were removed from analysis, and individuals for whom any of the dependent variables were greater than 3 standard deviations

from the overall participant mean were also removed from analysis. This resulted in the removal of 57 trials (1.14% of trials) which were outliers at the first response and 91 trials (1.82% of trials) which were outliers at the second response. After removing these trials, one participant was removed for being an outlier with respect to Answer Change, leaving 103 participants. Eight further participants were removed at random to bring the sample to the preregistered number, and all analyses were carried out with the remaining 95 participants.

Preregistered Analyses. Again, there was no statistically significant effect of delay on FOR ($M_D = 4.39$, $SD_D = 0.98$; $M_{ND} = 4.50$, $SD_{ND} = 1.03$; Figure 13), $t(94) = 1.90$, $p = .060$, $d = 0.20$, $BF_{10} = 1.35$, although the p-value is approximately near the chosen alpha level of $p = .05$ and the Bayesian analysis actually suggests weak evidence for an effect. There is, however, a clear effect on Answer Change (Figure 14), such that Answer Change was higher when there was a delay ($M = 29.04\%$, $SD = 19.01\%$) than when there was no delay ($M = 17.58\%$, $SD = 11.63\%$), $t(94) = 5.74$, $p < .001$, $d = 0.59$, $BF_{10} = 73.994$. There was no statistically significant main effect of the delay manipulation on Rethinking Time[5] (including only post-delay time), $t(94) = 1.66$, $p = .100$, $d = 0.17$, $BF_{01} = 1.03$.

[5] Although the preregistration for this experiment includes analysis of Rethinking Time, I maintain that the delay manipulation makes any such analysis difficult to interpret. As such, it is included here but consistent with Experiment 2a, it is not a main feature of my conclusion based on these analyses.

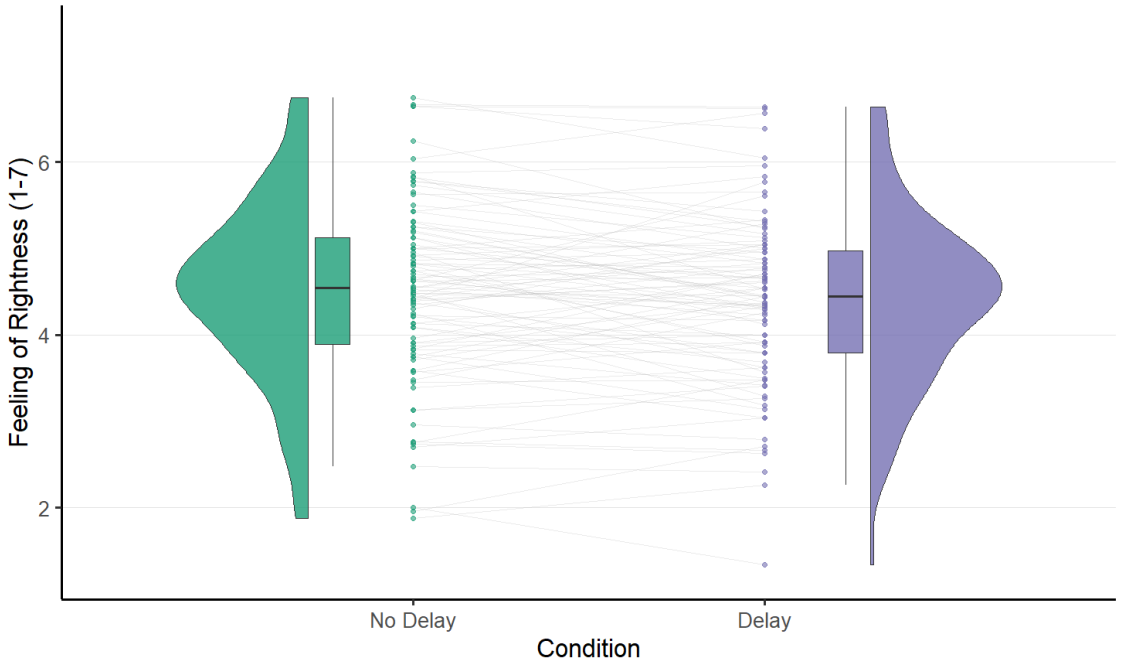


Figure 11. Feeling of Rightness (FOR) by Delay condition. Each pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles.

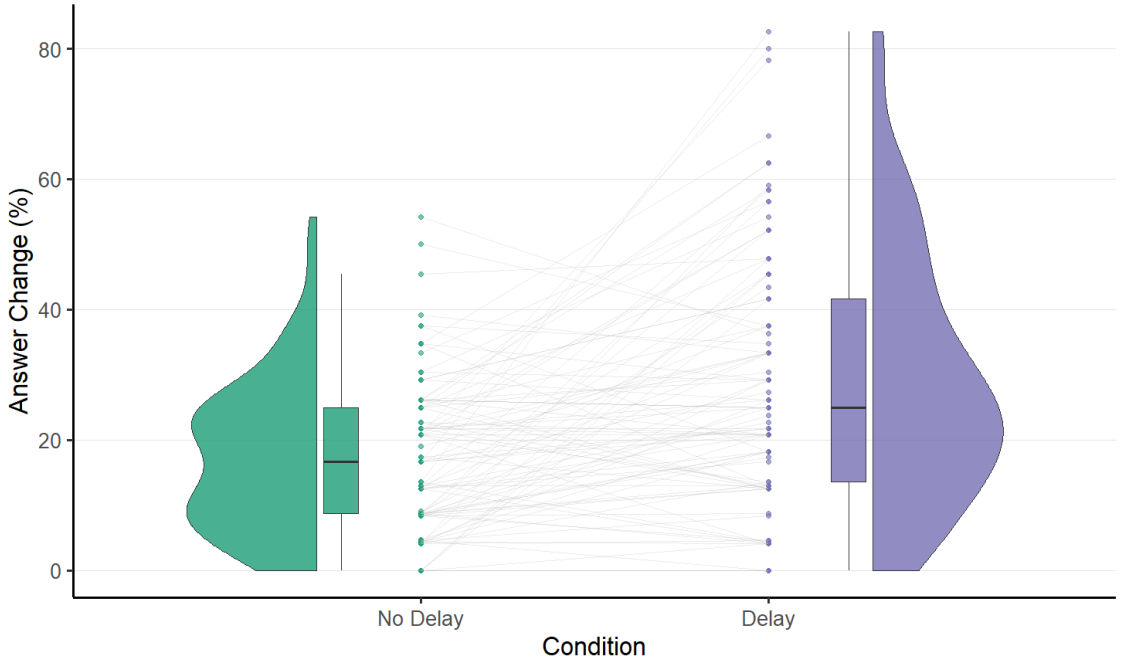


Figure 12. Answer Change by Delay condition. Each pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles.

Exploratory Analysis. The average within-subject correlation between FOR and Rethinking Time (r = -.34) was statistically significant, $t(94) = 14.08$, $p < .001$, $d = 1.44$, $BF_{10} = 1.98 \times 10^{21}$. The average within-subjects correlation between Response Time and FORs (r = -.23) was also statistically significant, $t(94) = 9.02$, $p < .001$, $d = 0.93$, $BF_{10} = 1.41 \times 10^{11}$. FORs were also lower for answers that were changed (M = 3.58, SD = 1.10) than those that were not changed (M = 4.65, SD = 0.7), $t(92) = 10.84$, $p < .001$, $d = 1.12$, $BF_{10} = 5.47 \times 10^{14}$. Note that two participants were removed from this analysis for never having changed an answer from their first response to their second. Item characteristics once again predictably impact FOR judgments, as FORs were higher when syllogisms were valid (M = 4.60, SD = 1.01) as opposed to invalid (M = 4.29, SD = 0.99), $t(94) = 5.92$, $p < .001$, $d = 0.61$, $BF_{10} = 149{,}586$. They were higher for syllogisms that were conditional (M = 4.52, SD = 1.03) than syllogisms that were categorical (M = 4.37, SD = 0.98), $t(94) = 2.70$, $p = .008$, $d = 0.28$, $BF_{10} = 5.46$. A Premise Believability x Conclusion Believability ANOVA revealed that both Premise Believability, $F(1, 94) = 37.35$, $p < .001$, $\eta^2_g = .028$, $BF_{10} = 2.46 \times 10^6$ and Conclusion Believability, $F(1, 94) = 10.10$, $p < .001$, $\eta^2_g = .008$, $BF_{10} = 13.60$ impacted FORs such that believable premises and conclusions were associated with higher FORs. This time, the Premise x Conclusion Believability interaction was statistically significant, $F(1, 94) = 15.23$, $p < .001$, $\eta^2_g = .009$, $BF_{10} = 52.18$, such that the effect was stronger when both premises and conclusions were believable.

2.5.3 Discussion

In Experiment 2b, I provide conclusive evidence that a delay manipulation targeting deliberation does not bleed into FOR judgments. This is strong evidence that, if there is a causal link between confidence and deliberation, it is specifically confidence that is the cause and deliberation that is the effect. In other words, these data rule out alternative models wherein the

processes underlying confidence and deliberation bidirectionally affect one another. Because the manipulation in Experiment 2a and 2b is similar in nature to the one in Experiments 1a and 1b, it is also unlikely that the process in Experiments 2a and 2b are qualitatively distinct from those in prior experiments. The most parsimonious explanation of the results of the four enclosed experiments largely aligns with the most face-valid interpretation of meta-reasoning processes to date: individuals monitor reasoning processes for cues informative to the extent that an item has been solved correctly, and they use the resulting sense of confidence to drive the relevant behaviour, in this case deliberation.

2.6 Chapter Review and Discussion

In the above four experiments, I provide perhaps the strongest evidence to date that the meta-reasoning process is causal in nature. This is a conclusion that has been widely presupposed by those studying metacognitive processes in reasoning; indeed, it is by far the simplest account of the relationship between confidence and deliberation. However, evidence of causality has been elusive, in part because of the nature of reasoning as a complex, higher-order process. In other domains, like perceptual decision-making, theoretical ground has been covered by neuroimaging (Boldt & Gilbert, 2022; Botvinick et al., 2001) and/or computational modelling (Alexander & Brown, 2010; Holroyd et al., 2005; Steinhauser et al., 2008) of the processes that perceive and translate visual information into a format that is useable by a control system. These models often contain causal components, explicitly or implicitly. This type of investigation is difficult to translate into fields like reasoning because of the increase in complexity and the decrease in ability to experimentally control individuals' processing of target stimuli. As such, I sought to address a specific set of what I consider the most plausible non-causal models in reasoning: those

where characteristics of object-level representations give rise to both confidence and deliberation.

Deliberation, however, can be strange to operationalize. Some view a lack of deliberation as the most consequential inter-individual failure of decision-making (Pennycook, 2023). In the current paradigm, deliberation is usually operationalized as Rethinking Time and Answer Change. However, in several places in the current set of experiments, these two variables do not align with respect to the effects of the chosen manipulations. In Experiment 1b, for example, there is a statistically significant effect observed on Rethinking Time but not on Answer Change. Why might that be? One reason might be the dichotomous nature of Answer Change as compared to the continuous nature of Rethinking Time. It is hard to argue that a tendency to spend more time thinking about a problem does not involve deliberation, but does every instance of answer change? Are there instances where deliberation is activated that *do not* result in answer change? To that question, the answer seems a definitive 'yes.' Many models allow for the possibility of the activation of effortful thinking resulting in rationalization of intuitive responses (Handley & Trippas, 2015; Pennycook, Fugelsang, & Koheler, 2015). As such, I conceptualize deliberation as a necessary but not sufficient condition for answer change; one cannot intuitively override an intuitive response, but one can deliberatively affirm an intuitive response. For this reason, trials on which responses were changed involve deliberative thinking and some trials where answers do not change involve deliberative thinking. For my purposes, this indicates that analyses of answer change likely underestimate effects on deliberation.

The pattern of results observed on deliberation is generally consistent with causal models and inconsistent with the specific class of alternative models described above. Manipulations of confidence produced (directionally opposite, because they are negatively related) approximately

equal effects on deliberation. That is, one component hypothesis of a causal model would be that it is possible to manipulate the causal variable and observe an effect on the outcome variable. This is clearly the case in Experiments 1a and 1b. Another component hypothesis is that a similar manipulation of deliberation *does not* produce an effect on confidence – if it is indeed the case that deliberation can be causally attributed to confidence. This is clearly consistent with the pattern of results shown in Experiments 2a and 2b. Most noncausal models are directly contradictory to this evidence, but not all of them. There remains at least one case where confidence and reflection both arise from characteristics of object-level processing. Imagine that some combination of feature characteristics inform both confidence and deliberation, but that deliberation is also based on a combination of processing factors (speed, etc.). It may be the case that the manipulations in Experiments 1a and 1b tapped into the shared space, and the manipulation in Experiment 2a and 2b tapped into the space uniquely predictive of deliberation. In that case, there is no causality necessary and the results could exactly mirror what I have observed. However, a model like this is made less likely by the fact that the manipulation across all four experiments was very similar. In other words, there are possible noncausal models remaining that can explain meta-reasoning processes and the results of the current four experiments, but they are very narrow. I can make no other reasonable conclusion: confidence causes deliberation.

Chapter 3: Refining an Item-Based Model of Confidence in Reasoning

3.1 Introduction

In Chapter 2, I provided the strongest evidence to date that monitoring and control processes in reasoning are causally linked. Particularly, control processes arise because of characteristics of the output of monitoring processes. With this in mind, one way empirical investigations of meta-reasoning can establish effects on control processes is *through* monitoring processes. That is, it supports the possibility that efforts to debias or otherwise improve reasoning by increasing the amount or effectiveness of deliberation might do so by manipulating the reasoning environment such that confidence is either globally reduced (e.g., Meyers et al., 2023), or better calibrated with performance (e.g., de Bruin et al., 2017; Prowse Turner & Thompson, 2009; Rawson et al., 2011; Sidi et al., 2017). Similar efforts have been made in reasoning, though usually with the intention of manipulating fluency (Alter, 2013; Alter et al., 2013; Alter et al., 2007, but see Meyer et al., 2015, and Thompson et al., 2013) for the purpose of reducing confidence and thus promoting deliberative processing. However, fluency is but one of the predictors of confidence and thus but one of the plausible targets for impacting confidence.

Recent focus, for example, has been on understanding the dynamic relationship between item characteristics and confidence. Obviously, the nature of any given task constrains the characteristics items can have, and thus the relationship between items' characteristics and confidence. Possibly task-general, however, is the way that cues are integrated to form a sense of confidence, a topic that has been long studied. Original demonstrations of cue integration in metacognition can be attributed to Koriat's (2012) Self-Consistency Model, which holds that subjective confidence is a product of the agreement of the corpus of cues in a problem with each other. Tasks where the cues for a given response are in general alignment will be associated with high subjective confidence; those where the cues generally diverge will be associated with low

confidence. Crucially, this paradigm provides insight into why some problems will be solved incorrectly with high confidence: cues generally align in cuing the incorrect response (Koriat, 2012). Koriat demonstrated that this pattern holds across a number of tasks, including word matching (Koriat, 1976), general knowledge questions (Koriat, 2008), perceptual judgments, social judgments, and categorization judgments (Koriat, 2012).

More recently, some researchers have utilised this approach to understand monitoring processes in reasoning. For example, Bajšanski and colleagues (2019) demonstrated that syllogistic reasoning follows a similar pattern: consensually correct and consensually incorrect syllogisms are both associated with high subjective confidence judgments, while low confidence is reserved for those items with lower consensus (the proportion of respondents who provide the same response to that item). The given account for this pattern is that problems whose cues generally align are those that are likeliest to have a large relative difference in evidence for each potential option at the time of response. Regardless of the speed or order in which one processes the available cues, the dominant response will be preferred more often and by a wider margin than the weak response. Kosourikhina and Handley (2024) provided further evidence for this account by adding cues in favour of the weak response in high-consensus problems, and observing a resultant decrease in confidence; the evidence available for each possible response was closer with the provision of evidence for the weak response. These investigations suggest that not only are item characteristics a substantial driver of confidence, but the totality of item characteristics interact to form a cue base that collectively impact subjective confidence judgments.

These accounts of reasoning are reminiscent of a class of models in two alternative forced choice tasks called Evidence Accumulation Models (EAMs). EAMs of decision-making are

(normally computational) cognitive models that describe how decision processes unfold over time (see Evans & Wagenmakers, 2019, and Ratcliff et al., 2016 for reviews). Much of the discussion in the EAM domain is with the intention of arbitrating between them for broad or specific classes of cognitive process. That is not the intention here. However, there are insights provided by the common characteristics of these models which may be informative as we think about evidence accumulations as an example process behind reasoning and monitoring reasoning. For example, a tenet of EAMs is that evidence is accumulated until some evidentiary threshold is met, at which time a response will be issued (Desender et al., 2021; Usher & McLelland, 2001). Desender and colleagues (2021) also provided evidence that confidence judgments, at least in a perceptual decision-making task, are reflective of evidence accumulated at the time of a response (consistent with the cue consensus work listed above and my characterization of De Neys's (2023) model). Another interesting consideration vis-à-vis reasoning is the nature of the stopping rule. One class of models contain relative stopping rules; the response threshold is determined by the relative evidence for two options (e.g., Busemeyer & Townsend, 1993). Should one option outweigh another by some amount, that option will be chosen.

An interesting implication of applying this type of evidence accumulation framework to cue utilization accounts of metacognition is that items whose cues generally align should not only produce consensual responses across participants, but they should also be responded to more quickly across participants. This is consistent with anecdotal evidence from my own work, where an item's average response time is often the strongest predictor of confidence on a given trial. Indeed, an item's average response time across participants might be a more faithful indicator of within-item cue alignment than even that item's response consensus. Imagine the

following scenarios: a participant is evaluating a syllogism and all the evidence suggests the response should be "valid," but when they submit their response, they mis-click and choose "invalid." This will skew that item's response consensus but not its average response time. Likewise, imagine a syllogism contains approximately ambivalent cues, equally supporting each response. Participants might, *en masse*, exhaustively process that items cues and be relegated to a guess. However, it might be the case that guessing behaviour reduces the item to one chosen cue, or even that there is a random bias across participants toward one response or the other. In these cases, that item's average response time, but *not* its response consensus, will most closely track cue alignment. In this chapter, I integrate theory about evidence accumulation in refining a model of the impact of item characteristics on confidence in reasoning. Namely, I test the extent to which items' response consensus and also their average response time, in combination with known individual processing and item characteristics, predict confidence in reasoning.

3.2 Experiment 3

In Experiment 3, I try to refine an item-based model of confidence in decision-making. Specifically, I test a simple model that I intend to capture the degree to which item-based cues as a whole impact confidence judgments in the same reasoning task employed in Experiments 1 and 2. To this end, I test how predictive of confidence are two variables: response consensus and average RT for each item. I aim to answer three specific questions: i) are consensus and RT, while conceptually similar, discernable with respect to confidence, ii) can RT, when added to the existing model, increase the model's ability to predict confidence, and iii) does that model require both predictors, or can consensus be dropped? The end goal is to establish a small set of predictors that capture a reasonable amount of inter-item variance in confidence judgments and can be used in an experimental context to test whether there is a difference in item-cue use across

experimental conditions. To address these goals, analyses are conducted at two levels: the item level, to assess the degree to which my chosen predictors predict variations in average confidence for each item, and at the trial level, to establish their promise for use in a multi-trial experimental design.

3.2.1 Method

Participants. Forty undergraduate students from the University of Waterloo (82.5% female self-identified, M_{age} = 20.1) participated in exchange for partial course credit. This sample size was preregistered and is based on a simulation-based power analysis in the R package mixedpower (Kumle et al., 2021), using previously collected data, including the four experiments in Chapter 2 and two from published work (Stewart et al., 2023). This sample size allows us .967 power to detect the increase in model fit when adding average RT to a model already including response consensus for predicting confidence at the trial level (i.e., with a mixed-effects analysis). This was done with the aim of achieving .95 power after some number of trials were excluded.

Design. There are no manipulated independent variables in the present experiment (beyond the manipulations of syllogism type, believability and validity noted below). Rather, both predictor variables arise organically. They are: (1) response consensus and (2) average RT. The dependent variable is confidence, given on a 100-point slider scale. The predictor variables are both aggregated at the item level; that is, they are aggregated across the forty participants for each syllogism. Each syllogism will then have one value for each of the variables. Response consensus is defined as the absolute value of the difference between the average response given (a value between zero and one) and the midpoint (i.e., 0.5). Items where most responses are the same across participants have high consensus, and those where responses are highly variable

across participants have low consensus. Average RT is simply the average time with which participants responded to each individual syllogism.

Stimuli and Apparatus. Stimulus presentation and response recording was controlled by Qualtrics online survey software. The stimuli consisted of 48 individual syllogisms, which varied systematically in terms of their type (categorical vs. conditional), validity, and believability of their premises and conclusions. These stimuli have associated values that capture the degree to which their premises and conclusions are believable. The stimuli are the same as those in Chapter 2, taken from Solcz (2011; Appendix A).

Procedure. Participants were recruited through the University of Waterloo participant pool and redirected to Qualtrics to complete the experiment. Participants began by reading the consent and information form and indicating their desire to participate. They then began the trial phase. During the trial phase, participants saw a series of syllogisms in a randomly-presented order, to which they were to respond in either the affirmative (i.e., "yes, this syllogism is logically valid") or in the negative (i.e., "no, this syllogism is not logically valid"). The order of the response options "yes" and "no" (i.e., which of them appeared first in a vertical list) was randomized for each individual trial. Participant selected their preferred option with the mouse and, after making their selection, clicked a button below the response options to submit their response. There were no restrictions placed on the number of clicks participants could make before submitting their response. After solving each syllogism, participants indicated their level of confidence with regard to how likely they think it is that they answered the problem correctly. Participants did this for each of the 48 syllogisms. They then indicated their age and biological sex to complete participation.

3.2.2 Results

All data analysis and visualization was completed in the R programming language (R Core Team, 2017) with the assistance of add-on packages (see: data.table [Barrett et al., 2024]; lme4 [Bates et al., 2015], plyr [Wickham, 2011], sjPlot [Lüdecke, 2024]; stats [R Core Team, 2017], tidyverse [Wickham et al., 2018]). All analyses were carried out with an α of .05. Outliers were addressed at the individual and at the overall level. That is, each individual participant's trials with an RT Z-score of greater than 3 compared with that participant's other trials were removed from analysis, and individuals whose RT was a Z-score greater than three from the overall participant mean were also removed from analysis. This resulted in the removal of 37 trials (1.98% of trials), and two participants from analysis. All reported analyses are based on the resultant 38 participants.

Preregistered Analyses. The first set of preregistered analyses are conducted at the item level. As such, there are 48 observations for this set of analyses because that is the number of items in my set. These analyses are designed to test the degree to which response consensus and average RT reflect the totality of the cues in each syllogism by analyzing the degree to which they predict confidence. For the purposes of my main analyses, RT refers to the first click a participant makes when presented with a syllogism. Although not all trials involve only one click, and thus the first click used in the analyses may not actually correspond to the participant's answer *per se*, I have found in past work that first clicks usually perform best in analysis. For completeness and robustness, the analyses are reproduced using the other available timing mechanisms: last clicks and page submit times.

First, a model including both consensus and RT was compared to a model containing only a linear combination of the two variables to test the consensus-RT predictor subspace (i.e., the extent to which these predictors represent distinguishable predictors of confidence). The

model with each variable entered on its own was a better predictor of confidence than the model containing only a linear combination of the two variables, $F(2, 45) = 23.68$, $p < .001$. This suggests that, vis-à-vis confidence, these two variables predict non-fully-overlapping variance and therefore should be treated separately. Second, I compare a model with only consensus as a predictor to a model with both consensus and RT as predictors. The addition of RT significantly improves the model fit, $F(2, 45) = 8.77$, $p = .005$. This analysis also indicates that consensus must be retained as a predictor in the model, but I test this hypothesis directly by comparing a model with only RT as a predictor to a model with RT and consensus as predictors. Adding consensus significantly increases the model fit, $F(2, 45) = 21.31$, $p < .001$. The second of these analyses is the major contribution of the present study to the existing literature at the item level (i.e., I have added aggregate RT to past work already predicting confidence with consensus). The ultimate model I settle on, one with both consensus ($\beta = 0.53$, 95% CI [0.30, 0.76]) and aggregate RT ($\beta = -0.34$, 95% CI [-0.57, -0.11]), accounts for approximately 39% of the item-level variance in confidence ratings (Adjusted $R^2 = 0.392$).

The item-level analysis suggests that item-average RT and response consensus combine to form a good model of consensus and aggregate RT predicting confidence. Because there are a limited number of stimuli available, there are potential power concerns with these analyses. For example, if the effect of adding aggregate RT on model performance were smaller than could be reliably detected per my sensitivity analysis, there are no additional items that can be added. Conceptually, the cues present within a problem should be predictive of confidence at the aggregate level, but this should also be true at the trial level. Indeed, the logic behind item-aggregate confidence being a product of variables that capture cues as a unit operates through individual trials; as one is attempting to solve a syllogistic reasoning problem, one's confidence

on that problem should be sensitive to the cues present. If this is true in the aggregate, we will see item-level effects. However, perhaps the better test of the influence of aggregate cues on confidence is at the trial level, for this reason.

To solve the conceptual problem and the potential power problem, I repeat the above analyses using linear mixed effects models containing random intercepts nested within participants. This allows me to estimate each model with trial-level granularity. RT and consensus in the following analyses still refer to the item-aggregate values as the intention is for them to encapsulate the degree of agreement within the cues present in each item. As above, the model with each of the predictor variables achieved significantly better fit than the model with a linear combination of the two variables, χ^2 (1) = 39.20, $p < 0.001$, confirming that the two variables cannot be combined and must be treated separately. Also, like above, a model with consensus and RT entered as predictor variables was a significantly better fit than a model with consensus alone, χ^2 (1) = 13.81, $p = < .001$. Finally, a model with consensus and RT entered as predictor variables was compared to a model with only RT as a predictor. That model comparison was statistically significant, χ^2 (1) = 34.43, $p = < .001$. That is, consensus is required to be retained as a predictor in the model predicting confidence. The full mixed-effects model, like the item-level model above, performs reasonably well, explaining approximately 43% of the variance in confidence judgments at the trial level (Conditional R^2 = .425), however the fixed effects explain much less of the total variance than in the previous analysis (Marginal R^2 = .017). Once again, the analysis suggests consensus (β = 0.11, 95% CI [0.07, 0.14]) is a more important predictor than aggregate RT (β = -0.07, 95% CI [-0.10, -0.03]) in predicting confidence.

Exploratory Analyses. The above preregistered analyses expose a reasonable model containing two predictors I have chosen that allegedly capture items' composite cues, but no part

of the above analysis confirms that as true. That is, even if one accepts that these variables in tandem predict confidence to a reasonable degree, the process by which they do so remains unclear. One of two predictor variables has been established in the literature as a measure of an item's body of cues, and the other I argue is a product of the same process. To this end, the variables should be correlated with one another. There is no statistically significant correlation between an item's average response time and response consensus, $r(46) = -.074$, $p = .616$. This suggests perhaps these variables are not indicative of a similar process, but there are other potential reasons for this. Perhaps there is enough noise introduced in RTs by having participants respond over the internet via Qualtrics that this correlation is diminished. Experiment 4 in part addresses this possibility. Another possibility is that the mechanistic process behind the relations between aggregate RT, consensus, and confidence are more complex than proposed. I test this possibility by fitting a model including both i) the above predictors of confidence, and ii) the other characteristics of the items that correspond to cues individuals might use to inform their confidence judgments, namely Validity of the syllogism (i.e., whether it is logically valid), Believability of the syllogism (i.e., whether the syllogism is seen as true in reality), and syllogism Type (conditional or categorical). Notably, I do not include accuracy, as accuracy represents mostly a shift *between* an item's cues (specifically, toward Validity).

One prominent cue that impacts confidence (and accuracy) is Conflict between Believability and Validity, and as such I include the interaction between Validity and both premise and conclusion Believability. I also include all other second-order interactions for completeness. Table 1 shows the full model.

Table 1. Confidence predicted by item characteristics, Experiment 3 (unstandardized estimates).

Predictors	Estimates	CI	*p*
(Intercept)	77.67	40.03 – 115.31	**<.001**
Consensus	-4.81	-138.08 – 128.45	.944
RT	-1.10	-5.50 – 3.29	.623
Premise Believability	2.85	-8.11 – 13.81	.610
Conclusion Believability	-4.44	-11.60 – 2.73	.225
Validity [1]	10.92	-11.94 – 33.78	.349
Type [Conditional]	-24.07	-45.51 – -2.63	**.028**
Consensus × RT	2.71	-9.80 – 15.23	.671
Consensus × Premise Believability	-0.60	-13.30 – 12.10	.926
Consensus × Conclusion Believability	1.44	-13.58 – 16.46	.851
Consensus × Validity [1]	16.65	-23.82 – 57.11	.420
Consensus × Type [Conditional]	-34.33	-68.67 – 0.02	.050
RT × Premise Believability	-0.61	-1.99 – 0.77	.387
RT × Conclusion Believability	0.13	-0.71 – 0.97	.761
RT × Validity [1]	-1.48	-3.89 – 0.93	.228
RT × Type [Conditional]	2.59	-0.00 – 5.18	.050
Premise Believability × Conclusion Believability	0.81	-0.27 – 1.89	.143
Premise Believability × Validity [1]	-0.87	-3.24 – 1.50	.470
Premise Believability × Type [Conditional]	3.25	0.68 – 5.82	**.013**
Conclusion Believability × Validity [1]	0.86	-1.31 – 3.03	.437
Conclusion Believability × Type [Conditional]	0.76	-1.68 – 3.20	.541
Validity [1] × Type [Conditional]	1.73	-4.85 – 8.31	.606
Random Effects			
σ^2	428.81		
$\tau_{00\ ID}$	310.00		
ICC	0.42		
Observations	1780		
Marginal R^2 / Conditional R^2	.032 / .438		

There are two main pieces of evidence that suggest aggregate RT and consensus are capturing an item's cues. First, the addition of numerous individual cues to a model already including aggregate RT and consensus provides minimal improvement of the model's prediction of confidence (Conditional R^2 = .425 vs. Conditional R^2 = .438 with the addition of four new predictors and all second-order terms). Second, neither consensus and aggregate RT nor any of the added predictors (most of which normally predict confidence) are significant predictors in the full model. The latter piece is likely not because I no longer have the power to detect such relationships with the creation of a much more complex model, given there are significant effects of syllogism Type. Additionally, the estimate for Consensus predicting confidence becomes negative, suggesting substantial competition between it and the added predictors. As such, the conclusion I derive from this analysis is that consensus and aggregate RT are competing for variance with items' other characteristics, further evidence that they capture items' cues in the aggregate.

As mentioned above, I made the choice to treat the first click made by participants as their response time, but the data almost certainly includes some trials where participants subsequently switch their response. It is therefore likely the case that, at least on some trials, the operationalization of a response is imprecise. To test the extent to which this matters, I reproduce the preregistered item-level analyses here with response operationalized as last clicks (i.e., the final click made per response) and page submit times (i.e., participants submitting their response after having made their final choice). As long as i) the pattern generally holds, and ii) these models do not substantially outperform the above models, the choice of first clicks is justified. Using last clicks, the model with consensus and average RT was a better predictor of confidence than the model containing a linear combination of the two variables, $F(2, 45) = 21.85$, $p < .001$.

The addition of average RT as a predictor to a model containing only consensus significantly improves the model fit, $F(2, 45) = 7.79$, $p = .007$. Adding consensus as a predictor to a model containing only average RT also significantly increases the model fit, $F(2, 45) = 19.88$, $p < .001$. The final model with both consensus and average RT accounts for approximately 38% of the item-level variance in confidence ratings (Adjusted $R^2 = .385$), slightly worse than the original analysis. Using page submit times, the results are very similar; the model with consensus and average RT was a better predictor of confidence than the model containing a linear combination of the two variables, $F(2, 45) = 21.44$, $p < .001$. The addition of average RT as a predictor to a model containing only consensus significantly improves the model fit, $F(2, 45) = 8.07$, $p = .007$. Adding consensus as a predictor to a model containing only average RT also significantly increases the model fit, $F(2, 45) = 19.53$, $p < .001$. The final model with both consensus and average RT accounts for approximately 38% of the item-level variance in confidence ratings (Adjusted $R^2 = .384$).

3.2.3 Discussion

In Experiment 3, analyses revealed that i) response consensus and item-average RT have enough unique variance with respect to confidence that they need to be treated separately, ii) adding RT to a model already including consensus improves the fit of that model, and iii) consensus is still required to be retained as a predictor in said model. These two predictors predict a substantial amount of the variance in confidence judgments. That is, with a relatively simple model of two predictors and no higher-order terms, a reasonably strong predictive model can be created. Both standard linear modelling and linear mixed effects modelling were conducted and yielded identical results, establishing that processes shown to be at work at the item level translate to the trial level. However, these fixed effects explain noticeably less of the

variance in the mixed-effects analyses than in the item-level analyses, which is to be expected because the mixed-effects analyses include inter- and intra-individual variance that the item-level analyses do not. I also show that consensus and item-average RT absorb variance from a series of variables one could reasonably expect to predict confidence. This is evidence that the mechanism by which these two predictor variables predict confidence is by capturing individual cues in the aggregate. Exploratorily, I analyzed multiple different operationalizations of a response; the first click made, the last click made, and the page submit time. The evidence suggests the chosen timing mechanism, first clicks, provides the best approximation of a response in the given paradigm. However, there are almost certainly trials for which this is not the case. In addition, the order of the response options changed on a per-trial basis. This is by design to reduce the likelihood of participants responding randomly, but almost certainly introduces some degree of noise with respect to RTs. For these, in Experiment 4 I reproduce the experiment in an in-person setting in order to ensure that the timing of each response is precisely operationalized.

3.3 Experiment 4

Experiment 4 is largely an extension of Experiment 3, toward refining an item-based model of confidence in decision-making. Experiment 3 paints a relatively clear picture of i) the contributions of two identified predictors, namely item-aggregated response time and response consensus to each item, to reported confidence, and ii) the situation of those predictors within the broader process (i.e., that they are aggregating variance normally attributable to the individual characteristics of items). The results of Experiment 3 suggest that one could measure cue use as a whole by predicting confidence judgments with a relatively small set of predictors that I have identified. However, there are some reasons for caution in interpreting the results of Experiment

3. For example, I construe the first click a participant makes as the moment they decide, even though a fair number of trials involve more than one click. It is therefore plausible that the RT used for analysis does not always correspond to the moment a choice is made by a participant. Additionally, there remains the possibility that imprecision in response times introduced by online participation affects the results. Both of these issues can be solved by having participants complete the study in-person with a button-click response, which I do in Experiment 4.

3.3.1 Method

Participants. Forty-two undergraduate students from the University of Waterloo (63.2% female self-identified, M_{age} = 19.9) participated in exchange for partial course credit. A sample size of 40 was preregistered and is based on the same simulation-based power analysis as Experiment 3. Two or more participants will be excluded based on my exclusion criteria or at random to get to the preregistered sample size.

Design. There are once again no manipulated independent variables in the present experiment (beyond the variables associated with the syllogisms). The independent variables are response consensus and average RT, measured identically to Experiment 3. The dependent variable is confidence, given on a continuous slider scale with bounds 0 and 100. The predictor variables are both aggregated at the item level; that is, they are aggregated across all of the participants for each syllogism.

Stimuli and Apparatus. Stimulus presentation and response recording was conducted in Octave using Psychophysics Toolbox 3.0.15 (Brainard, 1997; Kleiner et al., 2007). The stimuli consisted of the same 48 syllogisms as in Chapter 2 and Experiment 3, taken from Solcz (2011).

Procedure. Participants were recruited through the University of Waterloo participant pool and completed the study in-person. Participants began by reading the consent and

information form and indicating their desire to participate. They then began the trial phase. During the trial phase, participants saw a series of syllogisms in a randomly-presented order, to which they were to respond in either the affirmative (i.e., “yes, this syllogism is logically valid”) or in the negative (i.e., “no, this syllogism is not logically valid”). These responses were given by pressing either the “1” key or the “3” key. The order of the response options (i.e., “yes” and “no”) was counterbalanced across participants such that each button corresponded equiprobably with each response option. After solving each syllogism, participants indicated their level of confidence with regard to how likely they think it is that they answered the problem correctly. Participants did this for each of the 48 syllogisms. They then indicated their age and biological sex to complete participation.

3.3.2 Results

All data analysis and visualization was once again completed in the R programming language with the assistance off add-on packages (see: data.table [Barrett et al., 2024]; lme4 [Bates et al., 2015], plyr [Wickham, 2011], sjPlot [Lüdecke, 2024]; stats [R Core Team, 2017], tidyverse [Wickham et al., 2018]). All analyses were carried out with an α of .05. Outliers were addressed at the individual and at the overall level by removing each individual participant’s trials with an RT Z-score of greater than 3 compared with that participant’s other trials, and individuals whose RT was a Z-score greater than three from the overall participant mean were also removed from analysis. This resulted in the removal of 32 trials (1.62% of trials), and one full participant from analysis. This leaves 42 participants, which is still more than the preregistered 40. Two more participants (those with the fewest number of trials with a response) were removed from analysis. All reported analyses are based on the resultant 40 participants.

Preregistered Analyses. Once again, the first set of analyses are conducted at the item level. Once again, these analyses are intended to assess the extent to which Consensus and average RT are 1) distinct from one another and 2) predictive of confidence. The model with consensus and average RT was a better predictor of confidence than the model containing only a linear combination of the two variables, $F(2, 45) = 15.97$, $p < .001$. The addition of average RT as a predictor to a model containing only consensus significantly improves the model fit, $F(2, 45) = 31.20$, $p < .001$. Adding consensus as a predictor to a model containing only average RT also significantly increases the model fit, $F(2, 45) = 13.28$, $p < .001$. The final model with both consensus ($\beta = 0.37$, 95% CI [0.17, 0.58]) and average RT ($\beta = -0.57$, 95% CI [-0.78, -0.36]), accounts for approximately 53% of the item-level variance in confidence ratings (Adjusted $R^2 = .529$). There are two apparent differences between the results of these analyses and the mirroring results in Experiment 3. First, the relative importance of the two predictors has switched; the stronger predictor is now average RT rather than consensus. Second, the model predicts noticeably more of the variance in confidence than it did in Experiment 3. These differences can both be explained by an overall decrease in noise (particularly in RTs) by having participants complete the study in-person as opposed to online.

As in Experiment 3, the item-level analyses are matched by multilevel analyses. As above, the model with each of the predictor variables achieved significantly better fit than the model with a linear combination of the two variables, $\chi^2 (1) = 20.24$, $p < 0.001$, confirming that the two predictors must be treated separately. A model with consensus and RT entered as predictor variables was a significantly better fit than a model with consensus alone, $\chi^2 (1) = 39.76$, $p = < .001$. Finally, a model with consensus and RT entered as predictors outperformed a model with only RT, $\chi^2 (1) = 16.85$, $p = < .001$. That is, consensus is required to be retained as a

predictor in the model predicting confidence. The full mixed-effects model, like the item-level model above, performs reasonably well, explaining approximately 43% of the variance in confidence judgments at the trial level (Conditional R^2 = .425), however the fixed effects perform much worse than in the previous analysis (Marginal R^2 = .017). Like the item-level analysis, this analysis suggests RT (β = -0.12, 95% CI [-0.16, -0.08]) is a more important predictor than consensus (β = 0.08, 95% CI [0.04, 0.12]).

Exploratory Analyses. Replicating the findings in Experiment 3, there is no statistically significant correlation between average RT and consensus, $r(46) = -.20$, $p = .177$, although it is descriptively larger than the correlation present in Experiment 3. Because this analysis occurs at the item, there are power concerns related to the limited nature of the item set I chose. That is, one could conceptually increase the number of observations enough to detect a correlation of that size, but it would require creating new items and re-piloting the set. A mixed-effects model with all available item characteristics – Validity, Premise Believability, Conclusion Believability, and Type – along with all available second-order interaction terms, predicting confidence was fit on the data. The results are shown in Table 2. Only average RT is retained as a significant predictor of confidence; consensus is eliminated as a significant predictor by the addition of various item characteristics. Characteristics that would normally be significantly predictive of confidence are erased via the inclusion of average RT and consensus. The addition of these characteristics to the model provides minimal benefit in terms of the proportion of variance in confidence judgments explained (an increase from Conditional R^2 = .356 to R^2 = .364). These results provide further evidence that consensus and RT are absorbing variance from these item characteristics.

Table 2. Confidence predicted by item characteristics, Experiment 4 (unstandardized estimates).

Predictors	Estimates	CI	*p*
(Intercept)	125.16	81.84 – 168.49	**<.001**
Consensus	68.62	-137.26 – 274.51	.513
RT	-5.25	-9.45 – -1.04	**.014**
Premise Believability	-5.02	-14.68 – 4.64	.308
Conclusion Believability	-6.61	-19.06 – 5.85	.298
Validity [1]	-20.99	-53.36 – 11.39	.204
Type [Conditional]	-1.13	-25.72 – 23.45	.928
Consensus × RT	-5.43	-20.59 – 9.74	.483
Consensus × Premise Believability	2.24	-10.91 – 15.38	.739
Consensus × Conclusion Believability	-4.96	-21.32 – 11.40	.552
Consensus × Validity [1]	25.37	-17.91 – 68.66	.250
Consensus × Type [Conditional]	15.79	-24.73 – 56.32	.445
RT × Premise Believability	0.35	-0.45 – 1.15	.395
RT × Conclusion Believability	0.65	-0.46 – 1.76	.251
RT × Validity [1]	1.63	-1.21 – 4.47	.260
RT × Type [Conditional]	0.88	-1.13 – 2.88	.391
Premise Believability × Conclusion Believability	0.44	-0.58 – 1.47	.396
Premise Believability × Validity [1]	-0.40	-2.75 – 1.95	.739
Premise Believability × Type [Conditional]	0.29	-2.07 – 2.64	.811
Conclusion Believability × Validity [1]	0.23	-2.65 – 3.10	.878
Conclusion Believability × Type [Conditional]	-2.16	-4.54 – 0.21	.074
Validity [1] × Type [Conditional]	-1.40	-8.42 – 5.62	.695
Random Effects			
σ^2	416.22		
$\tau_{00\ ID}$	215.72		
ICC	0.34		
Observations	1799		
Marginal R^2 / Conditional R^2	0.035 / 0.364		

3.3.3 Discussion

The results of Experiment 4 largely match the results of Experiment 3. Response consensus and item-average RT are both required to be entered separately into a model that predicts a substantial amount of variance in confidence judgments. These two predictors account for noticeably more of the variance in confidence judgments than in Experiment 3, which can probably be attributed to an overall reduction in noise – especially vis-à-vis RT – via participants completing the task in-person. This is true at the item level and using a mixed-effects analysis capturing data at the trial level. Experiment 4 also provides confirmatory evidence to the nature of the relationship between consensus and average RT and confidence, namely that they collide enough with the other individual cues present in a syllogism to render them no longer significant in a full model. One change from Experiment 3 is that average RT remains a significant predictor of confidence in this omnibus model in Experiment 4. Perhaps the simplest explanation for this is that the reduction in RT noise yields predicted variance over and above that which is shared with the other item characteristics.

3.4 Chapter Review and Discussion

There is tremendous potential benefit to our understanding of the metacognitive processes arising from decision-making and other cognitive processes from being able to study the corpus of cues present in a problem. Unfortunately, building a model of confidence from all known individual cues themselves is difficult because there can be, depending on the chosen task, many that it may require very large data sets to appropriately model them. Experimental restriction of the number of possible cues, oppositely, might yield particularly anti-ecological tasks. Not only is it difficult to model all known cues, but also there are an unknown number of unknown or potentially immeasurable cues at work as well that cannot appear as variables from

which to predict confidence. Recent work has tackled this problem head-on. Bajšanski and colleagues (2019), for example, offered an account that attached the response consensus present for an item to the consensus of the cues therein.

Here I forward the conclusion that, in addition to consensus, there is additional variance from the corpus of an item's cues that can be captured by adding the item's average response time which, when combined with consensus, form a relatively simple predictive model of confidence. The simple model on which I settle on predicts a substantial amount, especially given its simplicity, of the variance in confidence judgments at the item level. In an in-person study, it predicted over 40% of the inter-item variance in confidence using only two predictors. Much less, but still a statistically significant amount, of the variance in trial-to-trial confidence was explained by the model. However, the current investigation i) establishes that such a pattern detectably exists in trial-level granular data, and ii) establishes the relative size of that relational pattern. That is, these effects are large at the item-level, and detectable but small at the trial level. I also establish, at the trial level, that the addition of various known item characteristics does not improve upon the simple model I forward. A convenient prediction therefrom is that, even if some of the unknown characteristics of items that drive confidence were to become measurable, this simple model may already capture the variance attributable to those characteristics. That is, because this model containing second-order characteristics already encapsulates known first-order characteristics, it is a straightforward extension to assume that it captures other first-order characteristics as well.

The link between second-order characteristics and confidence judgments may be interesting and/or fruitful, but it is far from new. Treatment of confidence as a second-order judgment (i.e., as behaviour about behaviour) is a facet of many modern theories of

metacognition outside of the reasoning literature (e.g., Fleming, Dolan, et al., 2012), and there are interesting parallels between the current set of experiments and this literature. For example, I interpret my results as a suggestion that confidence is based largely on factors related to second-order characteristics of items. Unclear and unaddressed in this current line of work is whether these relationships are because confidence arises from second-order characteristics of metacognitive cues themselves (i.e., consensus and average RT), or because both second-order characteristics and confidence arise through some process from the first-order characteristics (e.g., believability, validity, etc.). However, it is at least plausible that confidence arises from first-order characteristics (Kepecs & Mainen, 2012). Unknown is the extent to which this extends to reasoning, but my results are consistent with a process that begins with some form of evidence accumulation, from which both second-order characteristics and confidence judgments arise. As I showed in Chapter 2, it is likely from these confidence judgments that deliberation arises (i.e., we can think of deliberation as a plausibly third-order behaviour).

One benefit of the current work is that it suggests one can study experimentally the extent to which individuals are using items' intrinsic cues when formulating their confidence judgments as opposed, for example, to experience-based cues, at least to the extent that such cues are non-overlapping. Currently, study of metacognitive monitoring in reasoning and decision-making is fixated on shifts between cues. For example, those who attend less to the believability of syllogisms are less prone to belief bias (Sá et al., 1999). Crucially, those who reliably detect conflict between cues are less susceptible to many errors of thoughtlessness (De Neys, 2014; Franssens & De Neys, 2009; Pennycook, Fugelsang, & Koehler, 2015). The current work provides an avenue to analyze the extent to which an item's totality of cues impact confidence. This presupposes that total cue use is not zero-sum (i.e., that any change in the use of an

individual cue is not matched by a commensurate and opposite shift in use of other cues), which is not explicit in most theories of cue use. However, it has been established, at least, that there are individual differences that predict the extent to which cues like validity – or item features mimicking validity – are processed, and that those differences are not offset by differences in, e.g., processing of believability (Bialek et al., 2020; Ghasemi et al., 2022). To conclude, it is likely the case that one can assess the extent to which confidence is a product of item-intrinsic cues by testing the extent to which consensus and item-aggregated RT are predictive of confidence. From there, one can compare this relationship across conditions where there might *a priori* be expected a difference in the extent to which confidence judgments are created from item-based cues, as I do in Chapter 4.

Chapter 4: On the Effects of Explicit Judgments of Confidence

4.1 Introduction

In Chapter 4, I address the focal point of the present thesis: what characteristics of the reasoning process can be attributed to the standard use of explicit confidence judgments? I test the effect of measuring confidence explicitly on monitoring and control processes in reasoning. More specifically: does the standard method of assessing confidence by explicit report alter monitoring and control processes involved in reasoning? This is, on its face, a tricky proposition: how does one examine relations between confidence and other variables *without measuring it*? Fortunately, researchers in the field of comparative cognition have studied metacognition in nonhuman animals (for a review, see Beran et al., 2014). Studying comparative metacognition is a similarly tricky proposition, in that nonhuman primates are incapable of giving explicit confidence judgments. For many of the same reasons that the target research question of this chapter requires behavioural operationalization of metacognition, comparative work in nonhumans does as well (Smith et al., 2014). The most promising paradigm (Beran et al., 2015; Smith et al., 2020), the "Go-When-You-Know" paradigm, has been used to examine confidence in other apes. In this paradigm, the participant (e.g., a chimpanzee) completes a primary cognitive task at one location and receives a reward at another. After each trial, preceding feedback, the primate participants could move towards a reward dispenser or stay put. If they didn't reach the dispenser in time, they would forfeit the reward. Moving early, before feedback, ensured they were in place if a reward was earned. The optimal strategy is to move only when correct, which requires some ability to assess one's performance on the task. This pattern (i.e., one that assumes some degree of metacognitive monitoring) is observed in chimpanzees (Beran et al., 2015) and, to a lesser degree, Capuchin Monkeys (Smith et al., 2020).

4.2 Experiment 5

In Experiment 5, I adapt this paradigm for use in humans (a considerably easier proposition) and compare it to a standard explicit condition to test monitoring and control effects of explicitly measuring confidence. This adaptation requires the preservation of the critical aspects of this paradigm that allow it to study metacognition, without a requirement for participants to report their confidence. I estimate these to be: i) instead of a confidence judgment, a choice made by participants that allows them to optimize their reward, ii) that this choice 'pays off' based on their performance, and iii) so that always behaving as if one provided a correct response is suboptimal, trials on which participants choose as if they were correct but were not incur a cost (in this case, opportunity cost). The main hypothesis in this chapter is that at least some of the controlled behaviour (herein, reflection) is driven by the explicit judgment, rather than the underlying representation of confidence. It is unlikely to be the case that reflection is driven entirely by the presence of an explicit judgment (a possibility initially ruled out by Thompson et al., 2011). A prediction therefrom is that confidence and deliberation are less closely-related in the adapted condition where no explicit judgment is made than in the normal, explicit confidence condition.

4.2.1 Method

Participants. Ninety-five undergraduate students from the University of Waterloo (68.8% female self-identified, M_{age} = 19.9) participated in exchange for partial course credit. A sample

size of $N = 80$[6] was preregistered based on an *a priori* power analysis using the package Superpower (Lakens & Caldwell, 2021). Sample data were simulated wherein the effect of confidence on reflection in the normal condition was equivalent to $d = 1.00$ (a translation of the observed effect in the experiments to this point) and the effect in the adapted condition was equivalent to $d = 0.25$, in the same direction. A 2 x 2 mixed ANOVA with 40 participants per between-subjects group and the prescribed effects yields approximately 90% power to detect an effect.

Design. There are two independent variables in the present experiment. First, participants were assigned to one of two experimental conditions (Explicit vs Implicit confidence judgments; see Procedure below). Second, comparisons are made using confidence (High vs. Low) as an independent variable. In the Implicit condition, this is given as a dichotomous variable, whereas in the Explicit condition, it is given via Likert scale and converted to a dichotomous variable. This is not ideal but is necessary to maintain as few degrees of difference between the "Go-When-You-Know" paradigm as used in animals and the one used here. The key dependent variables were Confidence (also an Independent Variable as described above), Rethinking Time (time to respond to the second iteration of each problem), and Answer Change (a binary variable indicating whether an individual's response changed from their first response to their second). The latter two variables are designed to capture deliberation.

[6] It is now standard practice for me to oversample my preregistered sample size by about 10% so that, if exclusions are warranted for any reason, it is unlikely that they will result in a sample that is smaller than preregistered (as is the case in some of the earlier experiments in this thesis). Because this experiment contains multiple mechanisms by which to reduce the number of valid trials recorded for each participant, it was oversampled by 10% (i.e., roughly 8 participants) *per condition*, hence the larger-than-normal difference between the collected sample and the analyzed sample.

Stimuli and Apparatus. Stimulus presentation and response recording was conducted in Octave using Psychophysics Toolbox 3.0.15 (Brainard, 1997; Kleiner et al., 2007). The stimuli consisted of the same 48 syllogisms as in Chapters 2 and 3, taken from Solcz (2011).

Procedure. Participants were recruited through the University of Waterloo participant pool and completed the study in-person. Participants began by reading the consent and information form and indicating their desire to participate. Participants were informed they would be participating in exchange for entries into a raffle to win one of two $25 Amazon gift cards, and that better performance would result in more entries into the raffle. Before beginning the to-be-analyzed trials, participants completed three practice syllogisms. They then began the trial phase. During the trial phase, participants saw a series of syllogisms in a randomly-presented order, to which they were to respond in either the affirmative (i.e., "yes, this syllogism is logically valid") or in the negative (i.e., "no, this syllogism is not logically valid"). These responses were given by pressing either the "1" key or the "3" key. Each trial involved responding to each syllogism twice, as in Chapter 2. The order of the response options (i.e., "yes" and "no") was counterbalanced across participants such that each button corresponded equiprobably with each response option. A counter was displayed at the top of the screen indicating the number of remaining trials. In this experiment, unlike previous experiments, participants were only given six seconds to respond to the first iteration of each syllogism, which was normally worth one raffle entry if answered correctly. This forced speeding is to prevent increased deliberation during the first response to each syllogism given the potentially increased motivation – to maximize their reward – of participants relative to prior experiments. After solving each syllogism, participants indicated their level of confidence regarding how likely they think it is that they answered the problem correctly. It is this confidence judgment that differed

across conditions. In the Explicit condition, participants were prompted with an adapted FOR judgment on a six-point Likert scale (Appendix C). In the Implicit condition, I tried to reproduce the "Go-When-You-Know" paradigm above for human participants. Participants indicated whether they would like the response they had just given (i.e., their *first* response to each syllogism) to be worth one raffle entry or five raffle entries if answered correctly. If they answered "five," they forfeited a future trial. In essence, optimal responding for the purpose of maximizing raffle entries requires one to "spend" a future trial when they are relatively certain they have answered a syllogism correct on the first iteration. This approximates the "Go-When-You-Know" paradigm because i) participants must strategically wager when they are correct in order to maximize raffle entries, ii) the success of this strategy depends on their performance, and iii) never wagering reduces individuals' entry accumulation potential. Entry maximization requires one to wager past some threshold of certainty[7] that one has answered correctly. In the Implicit condition, second responses were always worth five entries to incentivize participants to take their time to ensure they are correct, as was instructed in past experiments.

4.2.2 Results

All data analysis and visualization was completed in the R programming language with the assistance of add-on packages (see: BayesFactor [Morey & Rouder, 2024]; gghalves [Tiedemann, 2022]; lme4 [Bates et al., 2015]; lmerTest [Kuznetsova et al., 2017]; lsr [Navarro, 2015]; plyr [Wickham, 2011]; psych [Revelle, 2024]; sjPlot [Lüdecke, 2024]; tidyverse [Wickham et al., 2019]), and with data visualization guidance from van Langen (2020). All

[7] It is difficult to calculate optimal behaviour numerically here, as it depends on i) one's likelihood of answering the current trial correctly (for our purposes, confidence), but also ii) their likelihood of answering a future trial correctly, and iii) their then estimation that they will have answered that future problem correctly. In any case, thank you to an unnamed participant in the experiment, after whose participation we engaged in a lengthy discussion about this topic, concluding that there is some threshold of confidence past which it is optimal to wager a future trial to increase the value of the current one.

analyses were carried out with an α of .05. Outliers were addressed at the individual level by removing each individual participant's trials with a Rethinking Time Z-score of greater than 3 compared with that participant's other trials. Response Time (for the first response) is not similarly trimmed because it is perfectly bounded between zero and six seconds. This resulted in the removal of 48 trials (1.44% of trials). There are two mechanisms in the present experiment that can reduce a participant's number of trials: i) failing to provide a response before six seconds has elapsed, and ii) spending a future trial (in the Implicit condition only). As a result, there are variable numbers of trials per participant and considerably fewer overall trials in the Implicit condition. Because the preregistered analyses are aggregated at the individual level, it is likely the case that there will be substantial variance in the number of trials represented in each participant's data. To combat the noise that potentially arises from this, and to reduce the sample to be consistent with the preregistration, I excluded the participants from each condition with the fewest number of observations (eight participants from the Implicit condition and seven from the Explicit condition until there are $n = 40$ per between-subjects group). This approach is non-random because participants who almost always or almost never wager will be excluded, a problem that is addressed in part by the next experiment.

Preregistered Analyses. The first analyses test whether the explication of confidence affects the extent to which deliberation is a product of confidence. Because of the way confidence has been operationalized in this experiment (i.e., in both conditions as High Confidence and Low Confidence), the main tests of this hypothesis are 2 (Condition: Explicit vs. Implicit) x 2 (Confidence: High vs. Low) mixed ANOVAs with each of our measures of deliberation. On Rethinking Time (Figure 15), there was a statistically significant main effect of Confidence, $F(1, 78) = 69.02$, $p < .001$, $\eta^2_g = .108$, $BF_{10} = 3.28 \times 10^9$, such that Rethinking Time

was greater for Low Confidence trials than High Confidence Trials. There was, however, no statistically significant main effect of Condition, $F(1, 78) = 0.20$, $p = .654$, $\eta^2_g = .002$, $BF_{01} = 4.30$, nor a statistically significant interaction, $F(1, 78) = 0.06$, $p = .814$, $\eta^2_g = .002$, $BF_{01} = 4.93$. On Answer Change (Figure 16), there was an identical pattern: a statistically significant main effect of Confidence, $F(1, 78) = 69.64$, $p < .001$, $\eta^2_g = .277$, $BF_{10} = 5.49 \times 10^{11}$, but no statistically significant main effect of Condition, $F(1, 78) = 0.56$, $p = .457$, $\eta^2_g = .004$, $BF_{01} = 5.93$, nor a statistically significant interaction, $F(1, 78) = 0.02$, $p = .874$, $\eta^2_g < .001$, $BF_{01} = 5.96$. The results of these analyses suggest that i) the manipulation of confidence judgment type is having no effect on the relationship between confidence and deliberation, and ii) that dichotomizing confidence does not necessarily obscure real effects (given the clearly-detected effect of confidence on reflection). Another way to assess differences in the relationship between confidence and reflection (that does not require one to artificially dichotomize a scalar variable) is to compute within-person correlations between confidence and reflection and compare them across conditions. Here, there was no statistically significant difference in the correlation between confidence and Rethinking Time between the Explicit ($r = -.32$) and Implicit ($r = -.28$) conditions, $t(78) = 0.74$, $p = .460$, $d = 0.17$, $BF_{01} = 3.39$.

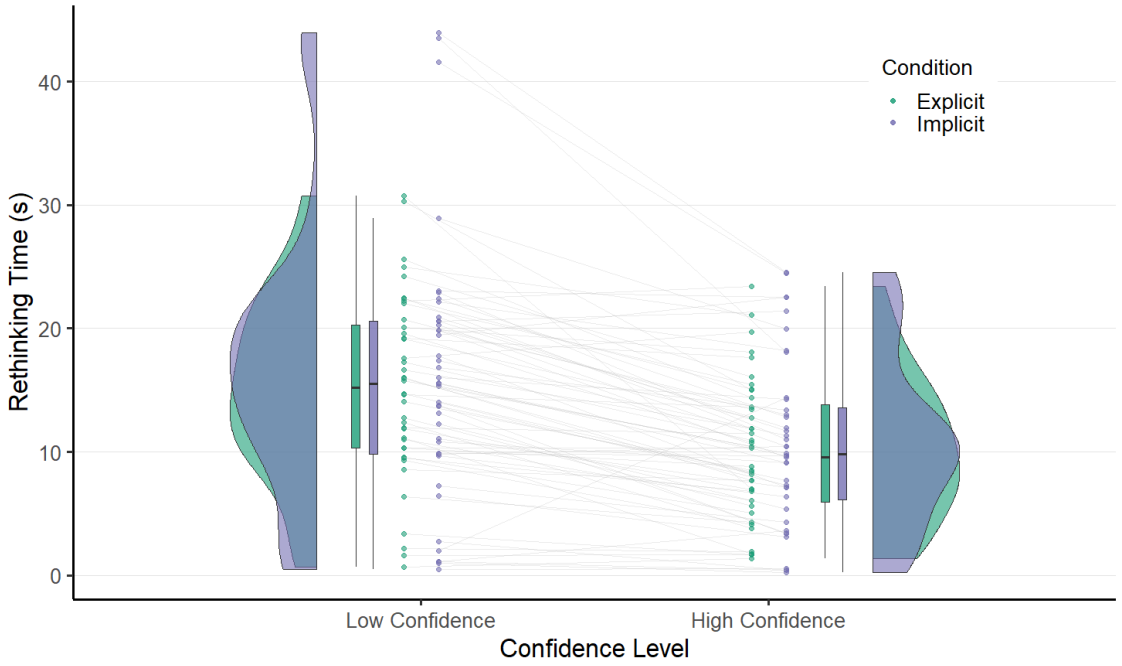


Figure 15. Rethinking Time by Condition and confidence level. Each connected pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles.

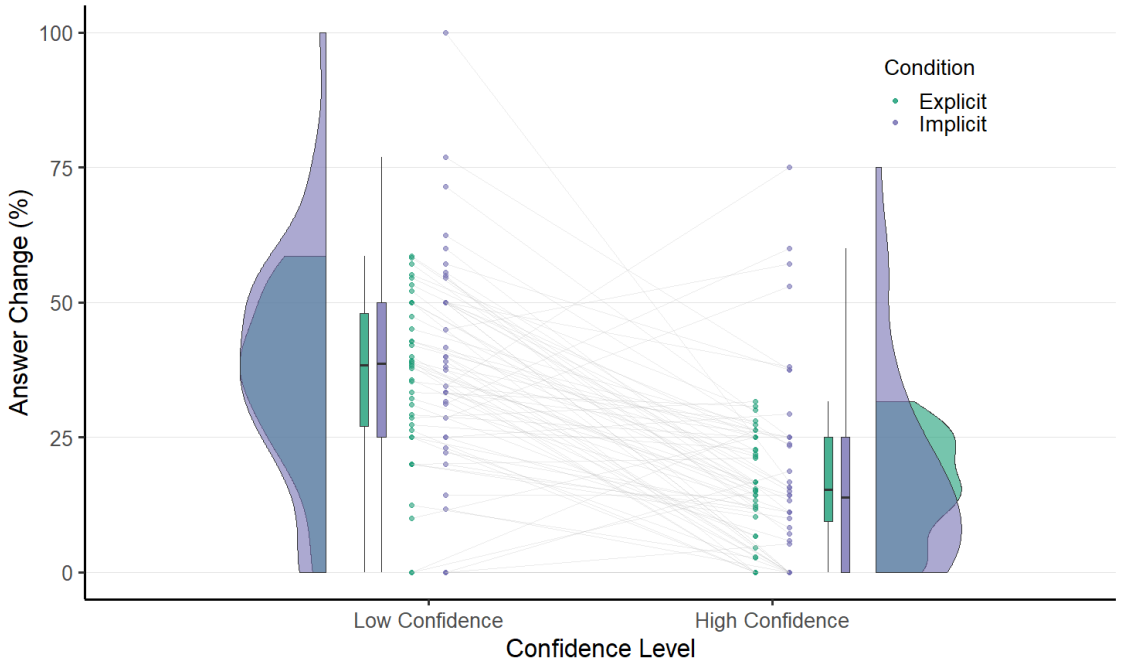


Figure 16. Answer Change by Condition and confidence level. Each connected pair of points represents an individual participant. Box plots display the mean and the first and third quartiles, and the whiskers extend 1.5 times the interquartile range beyond the first and third quartiles.

Exploratory Analyses. The main preregistered analysis aggregates data at the individual level, and as such, ignores noticeably different numbers of trials across participants. This leaves the possibility that noise, in participants with relatively few trials, drives results, especially given there are asymmetrically few trials in the Implicit condition. To combat this possibility, the analyses are mirrored with mixed-effects regressions with random intercepts nested within each participant. Unlike Chapter 3, these are not comparisons between nested models but rather a computation of the full model and interpretation of the individual estimates to best mirror the ANOVAs above. On Rethinking Time, there was once again a statistically significant main effect of Confidence, b = -5.49, t = 6.72, p < .001, but no statistically significant main effect of Condition, b = -0.35, t = 0.22, p = .824, nor a statistically significant Confidence x Condition interaction, b = -0.15, t = 0.15, p = .885. On Answer Change (because Answer Change is a binary variable, these are mixed-effects logistic regressions), there was a statistically significant main

effect of Confidence, b = -1.13, z = 7.17, p < .001, but no statistically significant main effect of Condition, b = 0.05, z = 0.29, p = .775, nor a statistically significant Confidence x Condition interaction, b = -0.07, z = 0.34, p = .736. The above analyses all confirm that the experimental manipulation of measuring confidence explicitly as normal or implicitly as adapted from primate research has i) no effect on deliberation overall, and more importantly, ii) no impact on the relationship between confidence and deliberation.

One conundrum in syllogisms and other reasoning tasks, however, is that it is often the case that deliberation is puzzlingly unrelated to accuracy, despite both accuracy and deliberation being touted as target behaviours (Stanovich & West, 2008; Thompson et al., 2011; Toplak et al., 2014; Evans, 2019). It is also the case that explicit prospective confidence judgments across other domains can alter or improve performance (e.g., JOL reactivity; Soderstrom et al., 2015; Double et al., 2018; Myers et al., 2020; but see Undorf et al., 2024 for a review). As such, I compare second response accuracy across conditions to test whether, even in the absence of an effect on deliberation, syllogism performance was affected by explicit confidence judgments (i.e., was explicit-judgment-reactive). There was no statistically significant difference in second response accuracy between the Explicit (M = 64.99%, SD = 14.89%) and Implicit (M = 62.58%, SD = 15.57%) conditions, $t(78) = 0.71$, p = .481, d = 0.16, BF_{01} = 3.46. Because the confidence judgment manipulation is between-subjects (i.e., participants only ever provide confidence judgments of one type), it could also theoretically be the case that participants anticipate the need to evaluate their initial response in order to give an explicit judgment and therefore any processing differences may take place during the initial response. There was, however, no statistically significant difference in accuracy for the first response ($M_{Explicit}$ = 55.82%, $SD_{Explicit}$ = 13.05%; $M_{Implicit}$ = 58.21%, $SD_{Implicit}$ = 14.26%), $t(78) = 0.78$, p = .437, d = 0.17, BF_{01} = 3.30.

The above preregistered and exploratory results provide relatively conclusive evidence that the difference in the requirement to explicitly report confidence does not produce any differences in controlled processing (i.e., deploying deliberative processing based on the assessed likelihood of being correct) nor overall deliberation.

Although the main preregistered analyses were designed to assess controlled processing (herein operationalized as deliberation), there remains the possibility, although less specific, that explicit confidence judgments may change *monitoring* processes, as opposed to controlled processes. That is, while I hypothesized that at least part of the relation between confidence and deliberation was due to the judgment itself (i.e., "I just said I was very confident, why would I expend more effort here?"), it is also plausible that confidence itself is produced differentially depending on what it is being produced *for* (e.g., to optimize one's own behaviour as in the Implicit condition, or to provide an explicit judgment). To test this possibility, I combine Condition with various predictors of confidence (many of which I established or verified in Chapter 3) to assess whether explicitly reporting confidence changes the extent to which certain item characteristics and/or processing factors impact confidence. Rather than compute a large, omnibus model, I break this analysis into coherent parts to address individual exploratory hypotheses. Every analysis below contains binary confidence (organically in the Implicit condition and converted in the Explicit condition) as the dependent variable. All are computed again with a random effects structure of only random intercepts by participant ID. To aid in model convergence, all continuous variables are standardized at the aggregate level (i.e., across both experimental conditions).

First, I compare a model with all item characteristics (Premise Believability, Conclusion Believability, Validity, and Type) entered as first-order terms to a model with all item

characteristics as well as a first-order term of Condition and interaction terms between Condition and all item characteristics. This analysis is designed to reveal the extent to which the differences in confidence type causes a shift in weight between these item-level characteristics in individuals' generation of confidence. The addition of Condition (and its associated interaction terms) does not significantly improve the model fit, $\chi^2(5) = 0.93$, $p = .968$. In addition, Condition did not significantly interact with any of the item characteristics, all ps > .603. In other words, the manipulation of confidence judgment type does not impact the relative translation of individual item cues into confidence.

Another plausible hypothesis is that the experimental manipulation alters the degree to which individuals use item-based or fluency-based cues as a whole. To test this possibility, I combine Condition with the aggregate-level cues identified in Chapter 3 (item consensus as well as item-average Response Time) as well as trial-level Response Time to examine the extent to which the requirement to give an explicit confidence judgment alters the makeup of confidence with respect to cues as a whole. There are statistically significant main effects of consensus, $b = 0.36$, $z = 4.81$, $p < .001$, and trial-level response time, $b = -0.40$, $z = 4.42$, $p < .001$, but not of item-average response time, $b = -0.01$, $z = 0.07$, $p = .942$. The addition of Condition (and its associated interaction terms) does not significantly improve the model fit, $\chi^2(4) = 2.56$, $p = .635$. Condition did not significantly interact with consensus, $b = 0.06$, $z = 0.61$, $p = .542$, item-average response time, $b = -0.01$, $z = 0.11$, $p = .910$, nor trial-level response time, $b = 0.14$, $z = 1.23$, $p = .219$.

Finally, I combine accuracy – for the first response – with Condition in a model predicting confidence to test whether explicit confidence judgments produce better- or worse-calibrated confidence judgments. The addition of Confidence (as a main effect and its interaction

with accuracy) does not significantly improve model fit, $\chi^2(2) = 4.20$, $p = .122$, and it does not significantly interact with accuracy, $b = 0.34$, $z = 1.91$, $p = .057$. However, there is a hint of an informative pattern of results here because i) the two above tests (especially the latter) approach statistical significance and ii) prior to the addition of Condition, accuracy is a strong predictor of confidence, $b = 0.41$, $z = 4.64$, $p < .001$, while the addition of Condition (and the interaction between accuracy and Condition) renders it not a statistically significant predictor of confidence, $b = 0.21$, $z = 1.51$, $p = .130$. Even though there is no statistically significant interaction between accuracy and Condition, this pattern of results is generally consistent with the existence of a relation between those two predictors. An examination of the predicted probabilities of a high confidence judgment (or a wager, in the Implicit condition) confirms that explicit confidence judgments are descriptively more sensitive to accuracy than wager decisions (Figure 17). These results, unlike the above results vis-à-vis controlled processing, paint a less-clear picture of effects of the experimental manipulation on *monitoring*. It is descriptively the case that the experimental manipulation impacts the degree to which individuals' confidence judgments are calibrated with their performance.

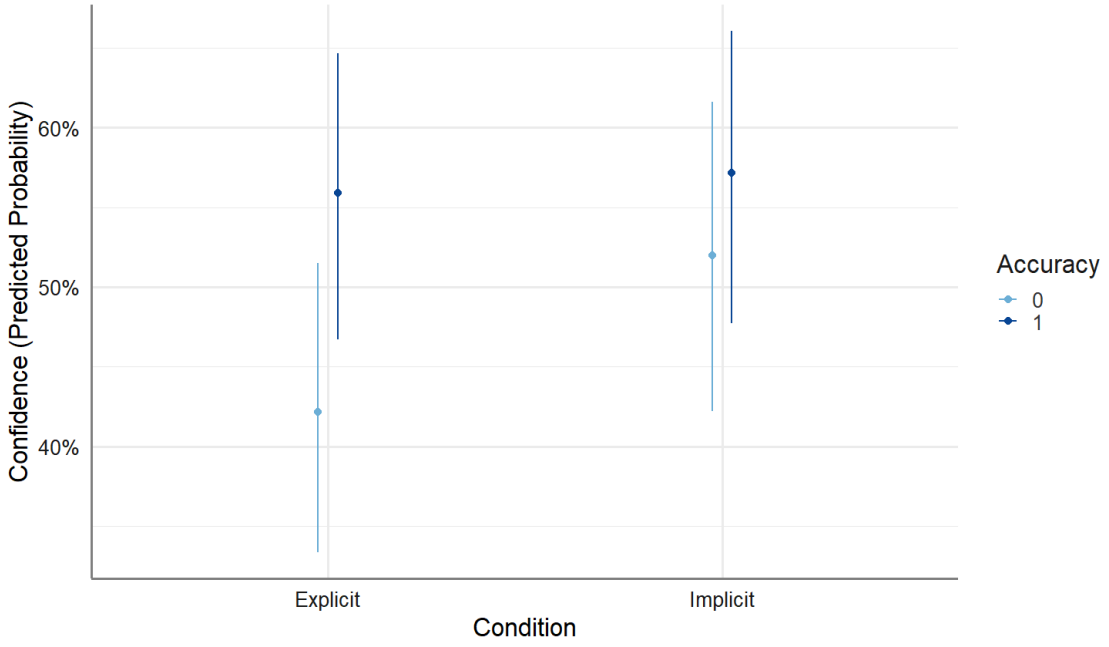


Figure 17. Predicted probabilities of a high confidence response by condition and accuracy. Error bars represent 95% confidence intervals.

4.2.3 Discussion

In Experiment 5, I provided a first test of the hypothesis that explicit confidence judgments, as often used in the literature, affect the metacognitive processes active in reasoning. Specifically, I hypothesized that at least part of the relationship between confidence and controlled behaviour, commonly operationalized as deliberative processing, is due to the requirement to give an explicit confidence judgment. In Experiment 5, I provide evidence against that hypothesis. In both a traditional group comparison and a mixed-effects regression analysis, confidence equally predicts deliberation across experimental conditions. That is, deliberation is presumptively caused by confidence as an underlying construct rather than some combination of confidence and the explicit judgment common in meta-reasoning investigations.

There is, however, another candidate mechanism by which requirements to explicitly represent confidence may alter meta-reasoning processes. One might hypothesize that individuals generate confidence differently depending on the ultimate purpose of the monitoring process, be that for the purpose of external report or in order to optimize one's own behaviour. There are a number of ways one can operationalize this hypothesis, but three possible conceptualizations are tested above: regarding confidence, it could plausibly be the case that response requirements alter i) the extent to which confidence is made up by individual cues, ii) the extent to which confidence is based on an overall sense of processing ease or difficulty, or iii) the extent to which confidence judgments align with actual performance. I find mixed evidence with respect to these hypotheses. There appears to be no shift in individual cue use between conditions, and neither does there appear to be a difference in use of overall processing cues (as measured by item response consensus, item-average RT, and trial-level RT). However, evidence for a difference in calibration as a function of condition is mixed. It may be the case that individuals' confidence responses – defined differently across conditions – are more sensitive to whether they responded correctly when those responses are explicit confidence judgments. Viewing Figure 17, it is clear that this pattern is driven by more frequent high confidence responses (i.e., wagers) in the Implicit condition. One possibility is that there is a real effect here that I do not have the statistical power to detect. Another possibility is that this pattern appears because of the treatment of the scalar outcome variable in the Explicit condition. Imagine, in a simple example, that the hypothetical wagering threshold for participants allocated to the Explicit condition would have been between '5' and '6' on the judgment scale. As long as there are some incorrect trials given '4' and '5' confidence ratings, bifurcation of their scalar judgment necessarily groups some hypothetical incorrectly-answered trials into the high confidence bin,

despite the fact they would not have resulted in a wager. Enough occurrences of this could produce the pattern seen in Figure 17 without a true difference (i.e., with *only* the differences in scale). Experiment 5 was designed with full awareness that the different response scales may be problematic, a limitation that I saw as necessary to adapt the paradigm with as few differences as possible from that used in nonhuman primates. Experiment 6 adapts one step further to allow for equality in response scale between the two experimental conditions.

4.3 Experiment 6

Experiment 6 is aimed toward answering the same hypotheses put forward in Experiment 5, but I adapt the design to solve some of the issues created by the design of Experiment 5. Specifically, the implicit confidence judgment (i.e., wagering) will no longer be a binary judgment, but rather will be made on a scale equivalent to that of the confidence judgment in the Explicit condition. This necessarily requires the design to move one step further away from the "Go-When-You-Know" paradigm introduced above. To ensure that the design of Experiment 6 allows conclusions about the same hypotheses as Experiment 5, it must meet the same three criteria set above, that the Implicit condition require i) a judgment made by participants to optimize their reward, ii) that this choice 'pays off' based on their performance, and iii) a cost incurred by not wagering when one responded correctly. To these ends, I replace the dichotomous option for participants to wager a future trial with a wager system that provides participants with seven raffle entries on each trial. They then have the option to bank those entries or wager them on their just-made response to a syllogism. In the case that they wager on a correct response, entries will multiply by five.

4.3.1 Method

Participants. Participants were 172 undergraduate students from the University of Waterloo (78.3% female self-identified, M_{age} = 19.7) who participated in exchange for partial course credit. A sample size of N = 150 was preregistered based on an *a priori* sensitivity analysis. Because no interaction between Condition and Confidence was detected with respect to deliberation, this effect, should it exist is likely very small. As a result, sample data were simulated wherein the effect of confidence on reflection in the normal condition was equivalent to d = 1.00 (approximately equivalent to all heretofore experiments) and the effect in the adapted condition was equivalent to d = 0.8 in the same direction (i.e., the difference in effect sizes between conditions is small, equivalent to d = 0.2). From there, the simulated data were resampled using the mixedpower (Kumle et al., 2018) R package, where it was determined for the simulated effect size that a sample size of N = 150 yielded 95% power in a mixed-effects regression analysis with random intercepts nested within participants to detect the interaction between confidence and condition.

Design. There are again two independent variables in the present experiment. First, participants were assigned to one of two experimental conditions (Explicit vs Implicit confidence judgments; see Procedure below). Second, confidence was assessed using – in both conditions – a seven-point scale. The key dependent variables were Confidence (also an Independent Variable as described above for some of the following analyses), Rethinking Time (time to respond to the second iteration of each problem), and Answer Change (a binary variable indicating whether an individual's response changed from their first response to their second). The latter two variables are designed to capture reflection.

Stimuli and Apparatus. Stimulus presentation and response recording was conducted in Octave using Psychophysics Toolbox 3.0.15 (Brainard, 1997; Kleiner et al., 2007). The stimuli consisted of the same 48 syllogisms as in Chapters 2 and 3 as well as the previous experiment, taken from Solcz (2011).

Procedure. Participants were recruited through the University of Waterloo participant pool and completed the study in-person. Participants began by reading the consent and information form and indicating their desire to participate. Participants were informed they would be participating in exchange for entries into a raffle to win one of two $25 Amazon gift cards, and that better performance would result in more entries into the raffle. Before beginning the to-be-analyzed trials, participants completed three practice syllogisms. They then began the trial phase. During the trial phase, participants saw a series of syllogisms in a randomly-presented order, to which they were to respond in either the affirmative (i.e., "yes, this syllogism is logically valid") or in the negative (i.e., "no, this syllogism is not logically valid"). These responses were given by pressing either the "1" key or the "3" key. Each trial involved responding to each syllogism twice, as in Chapter 2 and Experiment 5. The order of the response options (i.e., "yes" and "no") was counterbalanced across participants such that each button corresponded equiprobably with each response option. A counter was displayed at the top of the screen indicating the number of remaining trials. Again, participants were only given six seconds to respond to the first iteration of each syllogism, which was always (i.e., across both experimental conditions) worth one raffle entry if answered correctly. This forced speeding is to prevent increased deliberation during the first response to each syllogism given the potentially increased motivation of participants relative to prior experiments. Unlike Experiment 5, the counter always decreased by one per trial – as long as a response was provided within six

seconds – as there is no mechanism by which participants can reduce the number of future trials by wagering a future trial. After solving each syllogism, participants gave a judgment that was used to approximate confidence. This judgment varied across conditions. In the Explicit condition, participants were always awarded five raffle entries for each of their first and second responses to each syllogism. Their confidence judgment was unrelated to the number of raffle entries accrued and only reflected the extent to which they felt like their first response was correct. In the Implicit condition, participants were given seven raffle entries on every trial. They had the option to 'bank' these entries or wager between one and seven of them. Any banked entries would count as one entry at the end of the experiment, and any wagered entries would multiply by five if their initial response was correct, or zero if their initial response was incorrect. Once again, optimal responding for the purpose of maximizing one's number of raffle entries requires one to bank their entries when they responded incorrectly and wager their entries when they responded correctly. Despite changing the response scale from Experiment 5, this implicit confidence response retains the critical components because i) participants must strategically wager when they are correct in order to maximize raffle entries, ii) the success of this strategy depends on their performance, and iii) never wagering reduces individuals' entry accumulation potential. In the Implicit condition, second responses were once again always worth five entries to incentivize participants to take their time to ensure they are correct, as was instructed in past experiments.

4.3.2 Results

All data analysis and visualization was completed in the R programming language with the assistance of add-on packages (see: BayesFactor [Morey & Rouder, 2024]; gghalves [Tiedemann, 2022]; lme4 [Bates et al., 2015]; lmerTest [Kuznetsova et al., 2017]; lsr [Navarro,

2015]; plyr [Wickham, 2011]; psych [Revelle, 2024]; sjPlot [Lüdecke, 2024]; tidyverse [Wickham et al., 2019]), and with data visualization guidance from van Langen (2020). All analyses were carried out with an α of .05. Twenty-two participants (eleven from each experimental condition) were removed from analysis at random to reach the preregistered sample size of $N = 150$. Outliers were addressed at the individual level by removing each individual participant's trials with a Rethinking Time Z-score of greater than 3 compared with that participant's other trials. Response Time (for the first response) is not similarly trimmed because it is perfectly bounded between zero and six seconds. This resulted in the removal of 97 trials (1.56% of trials).

Preregistered Analyses. The main preregistered analyses are tests of the interaction between Condition and confidence predicting deliberation. Recall that I operationalize deliberation in two separate ways: Rethinking Time and Answer Change. These mixed-effects models are computed with a random-effect structure containing only random intercepts by participant. On Rethinking Time (Figure 18), there was a (at this point well-established) statistically significant main effect of Confidence, $b = -1.11$, $t = 12.23$, $p < .001$, but no statistically significant main effect of Condition, $b = -0.14$, $t = 0.13$, $p = .901$, nor a statistically significant Confidence x Condition interaction, $b = -0.13$, $t = 1.02$, $p = .310$. On Answer Change (Figure 19; a mixed-effects logistic regression because of the dichotomous nature of Answer Change), there was a statistically significant main effect of Confidence, $b = -0.28$, $z = 10.83$, $p < .001$, and this time both a statistically significant main effect of Condition, $b = 0.52$, $z = 3.29$, $p = .001$, and a statistically significant Confidence x Condition interaction, $b = -0.09$, $t = 2.48$, $p = .013$. Inspection of Figure 19 suggests this is due to a decrease in the tendency to change one's answers at the lowest confidence levels in the Implicit condition.

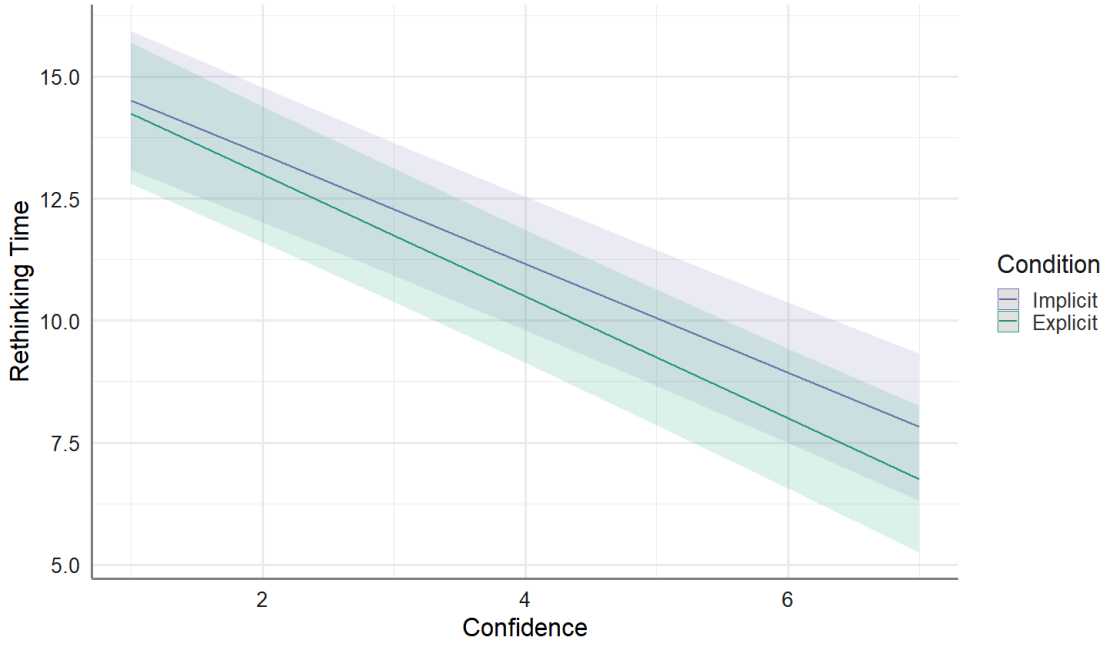


Figure 18. Predicted values of Rethinking Time by Condition and Confidence. Lines indicate predicted values, and shaded areas represent 95% confidence intervals.

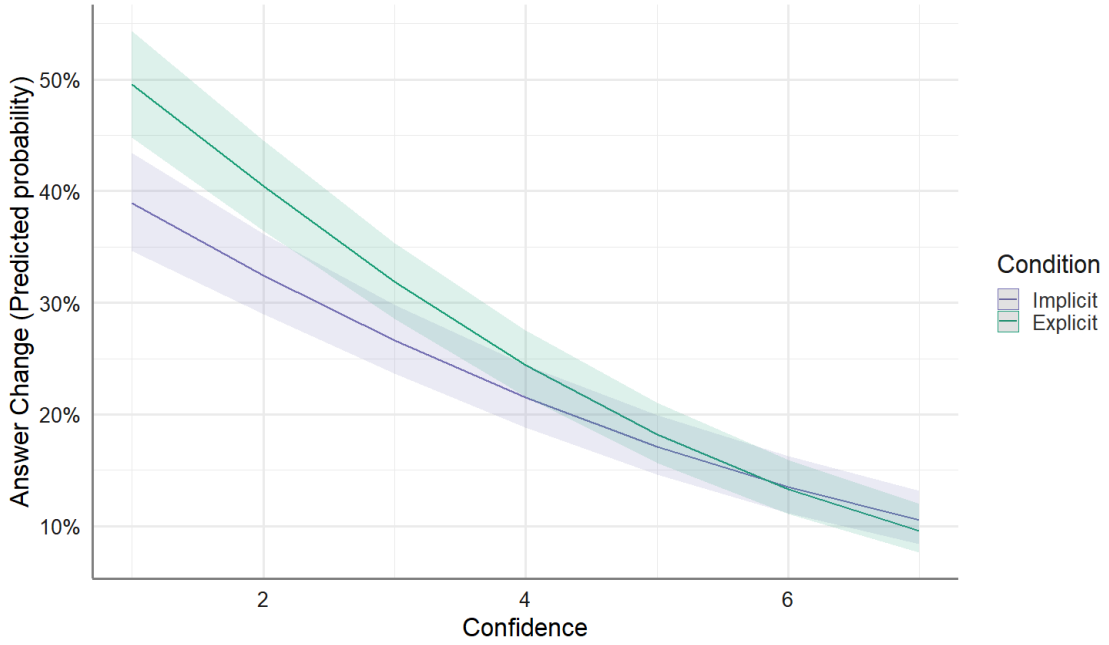


Figure 19. Predicted probabilities of Answer Change by Condition and Confidence. Lines indicate predicted values, and shaded areas represent 95% confidence intervals.

Further to analyse the extent to which the experimental manipulation of confidence judgment type influences control processes, analyses mirroring those of Experiment 5 were conducted to assess effects of the manipulation on *monitoring* processes. That is, does confidence judgment type influence the way(s) in which individuals generate a sense of confidence? There are many ways one can operationalize this umbrella hypothesis, but consistent with Experiment 5, I aim to assess the extent to which confidence judgments are based on i) individual cues, ii) an overall sense of processing ease or difficulty, and iii) the extent to which they align with actual performance.

First, item characteristics (Premise Believability, Conclusion Believability, Validity, and Type) were entered into a mixed-effects model in combination with Condition. The addition of Condition (and its associated interaction terms), like Experiment 5, does not significantly improve the model fit, $\chi^2(5) = 1.80$, $p = .876$. Condition did not significantly interact with any of the item characteristics, all ps > .659. Once again, Condition did not alter the extent to which confidence judgments – differently defined – arise from the individual cues present in the items I use. Second, I combine Condition with aggregate-level cues (item consensus as well as item-average Response Time) as well as trial-level Response Time to examine the extent to which the requirement to give an explicit confidence judgment alters the makeup of confidence with respect to overall processing cues. There were statistically significant main effects of consensus, $b = 0.24$, $t = 7.87$, $p < .001$, trial-level response time, $b = -0.21$, $t = 5.25$, $p < .001$, and (unlike Experiment 5) item-average response time, $b = -0.17$, $t = 3.78$, $p < .001$. Also unlike Experiment 5, the addition of Condition and its associated interaction terms significantly improved model fit, $\chi^2(4) = 29.14$, $p < .001$. Additionally, there was a statistically significant interaction between Condition and trial-level response time, $b = 0.29$, $t = 5.25$, $p < .001$ (Figure 20), such that the

relationship between RT and confidence is more strongly negative in the Implicit condition as compared to the Explicit condition (where it is noticeably positive). Finally, a model with both accuracy and Condition was fit predicting confidence judgments. The addition of Confidence does not significantly improve model fit, $\chi^2(2) = 2.41$, $p = .300$, and nor does it significantly interact with accuracy, $b = -0.09$, $t = 0.98$, $p = .329$ (Figure 21).

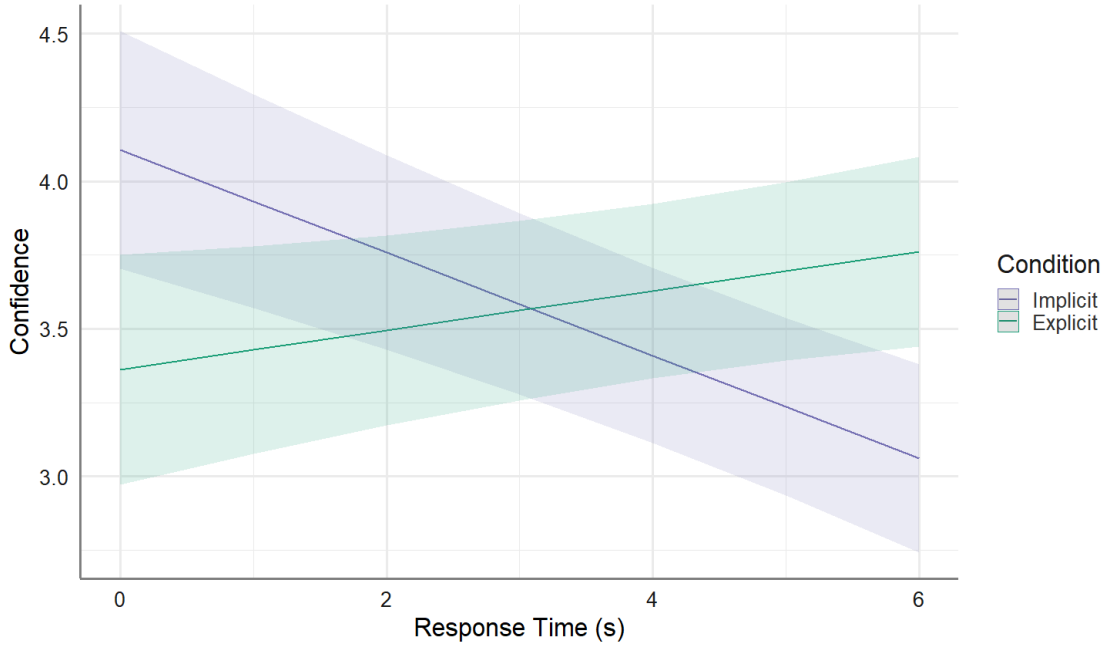


Figure 20. Predicted confidence judgments by condition and Response Time. Lines indicate predicted values, and shaded areas represent 95% confidence intervals.

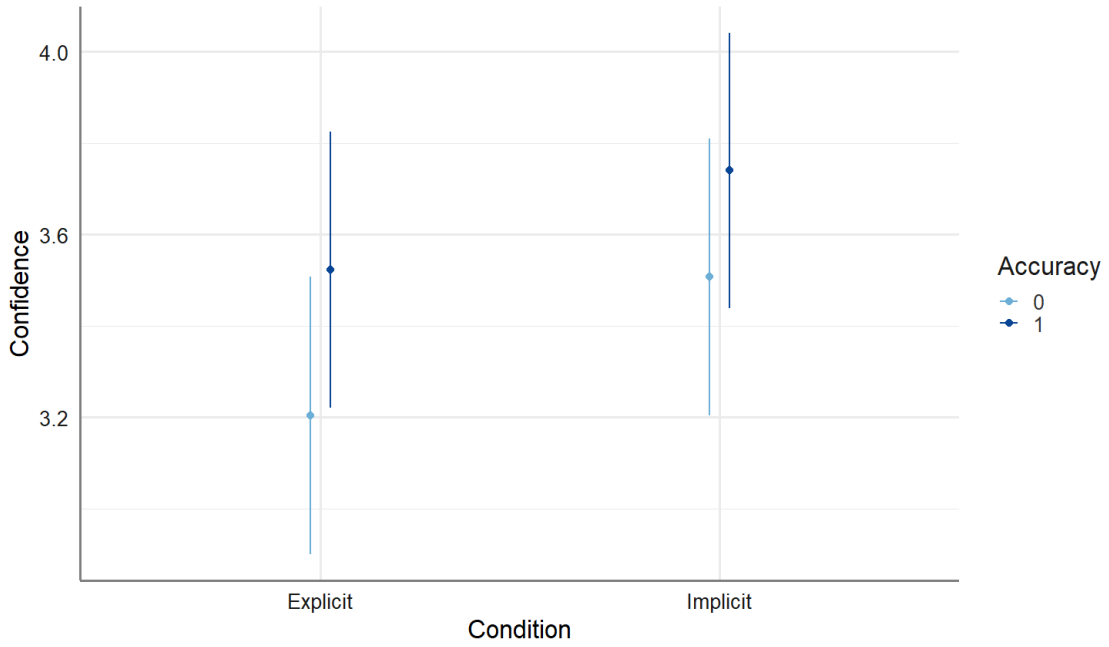


Figure 21. Predicted confidence judgment by condition and accuracy. Error bars represent 95% confidence intervals.

4.3.3 Discussion

In Experiment 6, I test the same set of hypotheses as in Experiment 5, but I do so in an experimental design that escapes the major problem(s) present in Experiment 5. The scale used in the Implicit condition in Experiment 6 is much more like the FOR scale to which it is compared than in Experiment 5. However, one could argue that it is inappropriate that it is treated *equally* to the FOR scale; this is a fair point. Is it reasonable to expect individuals to respond to a Likert-type scale in the same way they respond to a decision to wager some number of tokens? Probably not, but there is substantial evidence provided in Experiment 6 that they do. For example, many factors that have been demonstrated to predict confidence (normally on a similar scale to the one used here), do so equally to wager decisions. Whether or not wager decisions correspond to confidence, they are produced similarly. Perhaps the best evidence that wagering decisions are representative of something similar to confidence judgments is their relation to deliberation; each is approximately equally predictive of deliberation. In any case, it

must be that each measure is approximating similar constructs in similar ways, though I cannot rule out the possibility that the scales are used differently by participants. Relative to Experiment 5, however, the Implicit condition in Experiment 6 nevertheless allows cleaner conclusions vis-à-vis the effects that explicitly registering confidence has on monitoring and control processes in reasoning. I hypothesized that at least some of the relationship between confidence and controlled behaviour (herein defined as deliberative processing) was due to the confidence judgment itself. I mostly find evidence against this hypothesis. Confidence predicts Rethinking Time equally across experimental conditions; however, there are differences across conditions with respect to Answer Change, driven by a difference in deliberation at the lowest levels of confidence (Figure 19). Specifically, low confidence predicts increased Answer Change in the Explicit condition relative to the Implicit condition. There is no such pattern on Rethinking Time. There are three possibilities consistent with this pattern of data: i) there are real, small effects on controlled processing based on the requirement to evaluate confidence explicitly, ii) the analysis of Rethinking Time is detecting a true non-effect and the decrease in Answer Change for low-confidence Implicit trials is due to something other than decreased deliberation, or iii) as mentioned above, participants use these scales differently, to the extent perhaps that low responses in the Implicit condition do not exactly correspond to low FOR responses. Theoretically, Rethinking Time should be a more sensitive metric with respect to deliberation than Answer Change (even in circumstances where one deliberates, one will in at least some instances harness this deliberation to rationalize their intuitive response; Pennycook, Fugelsang, & Koheler, 2015; Wason & Evans, 1975). As such, it is unlikely to be the case that the Answer Change Analysis is capturing deliberation missed by the Rethinking Time analysis. In any case,

there is at most a very small effect of the experimental manipulation on control processes in Experiment 6.

In the other theoretical half of the presented analyses, I test the extent to which *monitoring* processes differ based on the ultimate judgment requirement (either a confidence judgment or a wager on one's performance). The pattern of effects here is somewhat clearer in that tests within the same hypothetical family are internally consistent. The experimental manipulation of judgment type evidently had no effect on the degree to which confidence responses were based on individual item characteristics such as item type, believability of the premises and conclusions, and validity, despite these characteristics being demonstrated to nominally affect confidence. Similarly, there was no effect on metacognitive calibration; participants' confidence responses equally discriminate correct from incorrect responses irrespective of the requirements of that confidence response. However, an interesting pattern arises in analysis of the relationship between confidence and variables designed to capture overall processing. Whilst the experimental manipulation of confidence judgment type did not impact the relation between an item's aggregate consensus response nor its average RT, trial-level Response Time was a significantly better predictor of wagering decisions than it was of confidence judgments. That is, an individual's ease or difficulty of processing an item informs wagering decisions to a greater extent than it does confidence judgments. Crucially, this cannot be attributed to the sense of certainty or uncertainty driven by the alignment or misalignment of the individual cues present within an item. That is, wagering on one's performance – as compared to providing explicit confidence judgments – seems to increase the extent to which the judgment relies on experiential (i.e., fluency-based) cues.

4.4 Chapter Review and Discussion

In Chapter 4, I tested the hypothesis that at least some of relationship between confidence and controlled behaviour in reasoning is driven by the requirement for participants to make an explicit judgment reflective of confidence itself. Experimental investigations regarding metacognition in reasoning normally include such a judgment, unlike how they might respond in the real world, to which this work is usually meant to extend. Rather, while real-world cognition may involve the evaluation of one's performance for the purpose of optimizing one's own behaviour, I contend it rarely if ever involves an explicit evaluation of one's confidence. To what extent, then, does the addition of such a judgment alter the metacognitive processes underlying reasoning? To provide an answer to this question, I adapted a paradigm from studies of metacognition in nonhuman primates to operationalize confidence implicitly for comparison against an explicit condition.

There was mixed evidence of an effect of explicit confidence judgments on the relationship between confidence and controlled processing (herein defined as deliberation and measured by Answer Change and Rethinking Time). In Experiment 5, confidence influenced deliberation – in each metric – equally across conditions; in Experiment 6, deliberation for low-confidence trials differed across condition. This difference is directionally similar, though not statistically significant, with respect to Rethinking Time. There are two implications from this finding. First, it is probably the case that wagering decisions approximate confidence to a reasonably similar degree that explicit confidence judgments approximate confidence. If wagering decisions were a poor proxy for confidence, one would not expect it to predict future behaviour to the same degree as direct judgments of confidence. Subsequently, this pattern of results supports the notion presented in past work (e.g., Thompson et al., 2011; Ackerman &

Thompson, 2017), and in Chapter 2, that it is confidence specifically driving controlled behaviour. In some ways, there is remarkable cohesion in the degree to which each operationalization of confidence predicts behaviour. On its face, deliberation on the basis of confidence is inherently rational; individuals can strategically allocate their limited cognitive resources where they are likely to pay dividends. Deliberation based on a wager vis-à-vis an item is considerably less rational, *unless* that wager corresponds roughly to one's confidence in responding correctly. The present results with respect to controlled behaviour are coherent with an account wherein wagering decisions are approximately analogous to confidence judgments, but either 1) with a small amount of error on low confidence trials, or 2) with the addition of a small number of non-metacognitive considerations in low-confidence trials.

It is worth considering what might drive a small decrease in deliberation for low confidence items in the Implicit condition. One pre-analysis concern was that deliberation in the Implicit condition might be impacted by extraneous factors such as risk aversion (Schneider & Lopes, 1986), ambiguity aversion for uncertain trials (Heath & Tversky, 1991), or other non-metacognitive, heuristic wagering strategies. For example, one may reasonably have predicted an *increase* in deliberation for low-confidence Implicit trials, whereby individuals effectively hedge (e.g., Shin & Ariely, 2004) their responses for particularly uncertain trials (i.e., provide one of each response to ensure they win at least some raffle entries when they are essentially guessing). It is important to note that I observe nothing like this in these data; indeed, the cross-condition equality of Rethinking Time in both experiments and Answer Change in Experiment 5 essentially preclude this possibility. Based on the present data, I cannot conclusively rule out the possibility that wagering decisions are a conceptually poor representation of confidence, but the difference

with respect to our variables of interest would need to nearly perfectly offset this difference. However unlikely, further research would be needed to conclusively rule out that possibility.

Likewise, the analyses of monitoring provide an account of wagering decisions as approximately equal to confidence judgments *per se*. For example, individual item characteristics such as Premise Believability, Conclusion Believability, Validity, and Type drive confidence judgments and wagering decisions equally. Similarly, aggregates of the cues present within an item, namely Consensus and Item-Average Response Time, are equally predictive of both types of judgment. This is not necessarily surprising; all these factors have been shown here and elsewhere (e.g., Stewart et al., 2023) to reliably impact confidence. Given wagering decisions should theoretically map closely to confidence (Persaud et al., 2007), these factors should relate equally to wager decisions as they do confidence. However, this is not true for all variables of analysis: trial-level Response Time is a substantially stronger predictor of wagering decisions than it is explicit confidence judgments. Perhaps unusually, however, it appears that this is driven by an effect on the Explicit condition, wherein the relation between RT and explicit confidence judgments is smaller (r = -.03) than even the same condition in Experiment 5 (r = -.14), despite the fact there is no apparent methodological difference between the two. As such, one might conclude that this is a random result, indicative of noise rather than some underlying truth. I urge caution against this conclusion, for two main reasons: First, Experiment 6 is cleaner, for the reasons outlaid above, and much more well-powered than Experiment 5. As such, ignoring the background theoretical expectations, the evidentiary value of Experiment 6 is greater than is Experiment 5. Second, the analysis supporting the conclusions made is not a raw comparison of these correlations: rather, this conclusion is based on a predictive model containing several variables. Illustratively, reference Figure 20: while the raw correlation

between RT and confidence judgments is unusually low in magnitude, but still negative, the isolated variance controlling for other variables suggests a positive predictive relationship between RT and confidence judgments. As such, even if the raw correlation between these two was of the expected magnitude, there would most likely still be an apparent difference on the basis of experimental condition in Experiment 6. Thus, while the nature of the evidence for the claim is not indisputable, I conclude that the adapted, Implicit condition increases the extent to which metacognitive judgments are based on experiential factors.

Chapter 5: General Discussion

5.1 Thesis Summary

Before I address the single most important conclusion of the present work, I will review the findings from Chapters 2 and 3 that allow full comprehension of the thesis. In Chapter 2, I provide the strongest evidence to date that metacognitive processes in reasoning are causal in nature; specifically, control processes (normally, deliberative processing) arise *because* of monitoring processes. To support this claim, I meet the standard set elsewhere: manipulations that affect confidence also affect controlled behaviour, even in the absence of effects on performance. I also go beyond this standard, by establishing that similar manipulations targeting controlled behaviour *do not* produce countereffects on the outputs of monitoring. Specifically, a delay in processing prior to a response inflates confidence and reduces deliberation, while a post-response delay (i.e., more time to think) promotes deliberation while leaving confidence unaffected. This, despite the fact that the delay comes before confidence judgments are made (and thus they could theoretically be affected by the manipulation). This series of four experiments aids in interpretation of both the present thesis as well as other investigations of metacognitive processing in reasoning as it allows one to infer effects on target behaviour *via* monitoring behaviour. Any effects of experimental manipulations, task instructions, individual differences, or item characteristics on confidence should also result in effects on deliberative processing, the primary instance of controlled processing in the field.

I also contribute to the study of these predictors of monitoring in Chapter 3, where I further common models of item characteristics on confidence. In reasoning research, there has been plenty of focus on characteristics of items and the effects they have on metacognition, and thus, reasoning performance. Some of this work has characterized the effects of individual cues

on metacognition (e.g., Brisson et al., 2018; Markovits et al., 2015; Stewart et al., 2023). Far more work has focused on interactions between subsets of the total set of cues (e.g., belief-logic conflict; De Neys et al., 2011; Pennycook et al., 2014; Thompson & Johnson, 2014). However, very recent investigations of meta-reasoning have taken a new approach to measuring and characterizing effects of the totality of cues within reasoning items. They have identified item-level variables (e.g., response consensus) that seem to logically correspond to the degree to which that item's cues align or misalign (e.g., Bajšanski et al., 2019; Kosourikhina & Handley, 2024), variables which predict a substantial amount of the variance in confidence judgments. Cue alignment measures predict even more of the variance in confidence than the previous field-level standard, conflict between logical validity and belief. I borrow from the logic of Evidence Accumulation Models to add another variable (item-average response time) which predicts unique variance in confidence judgments and substantially improves prediction of confidence when added to response consensus, both at the item and trial level.

Despite being meaningful theoretical contributions of their own, the primary value of Chapters 2 and 3, for the purpose of this thesis, is that they in tandem allow for far better interpretation and analysis of monitoring processes in reasoning. I use this to my advantage in Chapter 4 in order to address the main claim in the present thesis. In the two experiments in Chapter 4, I tested the possibility that researchers requiring participants to explicitly report their confidence in a reasoning task demonstrably alter the monitoring and control processes argued to underlie reasoning. These processes are argued to operate both i) in laboratory reasoning tasks where participants do not explicitly report their confidence (e.g., De Neys & Glumicic, 2008), and ii) in real-world settings where individuals almost certainly do not explicitly represent their confidence. I assessed this by adapting a very promising paradigm from the study of

metacognition in nonhuman primates, the "Go-When-You-Know" paradigm. This paradigm requires participants to monitor their performance and exert control on the task environment to maximize their rewards. In other words, it requires metacognition (Beran et al., 2015; Smith et al., 2020). I adapted this task for use in syllogistic reasoning, by requiring participants in an Implicit condition to retrospectively 'wager' on their performance, in lieu of explicit confidence judgments. In Experiment 5, this wager is in the form of a binary: participants can "spend," or "wager" a future trial to increase the value of the just-completed one. In Experiment 6, this wagering choice is changed to a continuous choice to wager up to seven "tokens" on each trial, and thus approximates a post-decision wagering task, the argued gold standard in behavioural operationalization of confidence (Kepecs & Mainen, 2012, but see Konstantinidis & Shanks, 2014).

In Chapter 4, I first demonstrate a small effect on controlled behaviour, wherein explicit confidence judgments are more closely related to deliberation than are wager choices. This effect appears only in Experiment 6, which has substantially better granularity, and while only statistically significant with respect to Answer Change (one of two measures of deliberation), it is directionally consistent across both measures. Specifically, answers are changed more for low-confidence trials in the Explicit judgment condition. More noticeably, the standard negative relationship between trial-level response time (when controlling for individual and aggregate item characteristics) was substantially stronger in the Implicit, wagering condition. That is, wagering decisions likely rely on experienced ease to a greater extent than do confidence judgments. The simple model developed in Chapter 3 provides the blueprint for this analysis, and the results of Chapter 2 suggest we can therefrom make inferences about deliberation. One would predict, then, that slower, less fluent responses should be associated with lower

confidence, and thus more deliberation, in scenarios where monitoring and control operate for the purpose of solely optimizing behaviour, rather than the provision of an explicit confidence judgment.

5.2 Situating these Results within the Meta-reasoning Framework

A question one might ask is whether the presented results are problematic for meta-reasoning accounts. After all, I justified the current set of studies by arguing that typical investigations may be altering the processes they allege to measure, by virtue of the inclusion of explicit confidence judgments. On the contrary, the tenets of meta-reasoning mostly hold firm. Despite a small increase in reflection for low confidence judgments (relative to small wagers), both conditions follow approximately the same pattern, supporting claims (e.g., Thompson et al., 2011) that control behaviour in meta-reasoning is based on confidence and that it is unlikely to be completely judgment-reactive. Likewise, item characteristics are translated equally into explicit and implicit confidence. Even the instance where there is the strongest effect lends support for standard claims: the finding that fluency predicts implicit confidence to a greater degree than explicit confidence judgments actually suggests that current accounts based on explicit judgments may be *underestimating* the size of the effects of fluency on confidence. In all, I can provide essentially no evidence that challenges major accounts.

One potential incongruity between the present work and the meta-reasoning literature is my characterization of confidence. Great care is commonly taken to differentiate between the many distinct types of metacognitive judgment available. Here, not only have I intentionally grouped them all together, but I directly equate Feeling of Rightness judgments with confidence. It is critical to understand that, with respect to the major hypothesis tests conducted, this difference is terminological rather than substantive. I do not mean that Feeling of Rightness

judgments as they appear in other investigations are tantamount to confidence judgments; rather, they are an instance that has often been used in conjunction with syllogistic reasoning and represents the body of judgment types I aimed to compare to an adapted condition. In essence, use of the term confidence judgments is meant to refer to the wider body of explicit, metacognitive evaluations of which FORs are simply a member. The thread between these judgments is that they correspond to some explicit evaluation of one's performance on some task. For my purposes, the most reasonable umbrella is confidence.

Experiment 6 reveals a somewhat substantial effect on the relationship between fluency and confidence as a result of the response requirement with regard to one's confidence. However, perhaps interestingly, there is no resultant difference in metacognitive calibration; despite relying more on their sense of ease while processing the problem initially, participants in the Implicit condition are no better at evaluating their performance. Is fluency a misleading cue for performance? This would be counter to claims made by others (e.g., Thompson et al., 2011), and yet consistent with demonstrations wherein fluency is experimentally manipulated independent of performance on a given task (Besken & Mulligan, 2013, 2014). So, what is one to believe? It is probably the case that in a standard experimental investigation, fluency is capturing multiple processes, some of which are useful and some of which are not. This raises the question: did the experimental manipulation here simply shift predictive weight vis-à-vis confidence onto only the components of fluency that are not useful? Or perhaps, are the normally useful portions of fluency instead captured by the novel way I modeled item characteristics? It is impossible to know for sure, but it is a reasonable avenue for future work to continue to investigate the complex relationships in metacognitive cue space. In any case, I characterize the manipulation as

shifting the weight toward a cue that is nondiagnostic (but notably not anti-diagnostic) with respect to performance.

The results of the current set of experiments provide both i) moderate evidence in favour of past claims about meta-reasoning, and ii) suggestions about some changes to the reasoning process as a result of the standard requirement for participants to evaluate their confidence explicitly. These effects can be informative with respect to the meta-reasoning framework. For example, they provide methodological suggestions about unintended effects the standard designs may be having on participants' reasoning. It may be a worthwhile endeavour to incorporate behavioural measures of confidence in the breadth of meta-reasoning research, which some researchers have already begun to do (e.g., Undorf et al., 2021). However, the implications go beyond methodology. One of the stated aims of meta-reasoning work is to explore the possibility of improving reasoning through metacognitive processes. The current work provides monitoring and control suggestions to this effect. First, it suggests that explicitly evaluating confidence may shift monitoring weight away from fluency or other experiential factors. As I have demonstrated, there are a number of instances (usually in intentionally-designed experimental settings) wherein fluency is at best nondiagnostic – and at worst misleading – regarding one's performance. The results of the present thesis (particularly Experiment 6) suggest that explicitly evaluating one's confidence may be a way to avoid such pitfalls. Second, the current work demonstrates small effects on controlled behaviour. One common approach is to attempt to improve monitoring accuracy, with the goal of allowing individuals to more efficiently allocate their resources (e.g., Batha & Carroll, 2007). A tandem to this kind of approach might be the strategic use of explicit confidence evaluations for the purpose of more closely tying controlled behaviour to the output of monitoring processes. For example, training or experimental manipulations that improve the

accuracy of metacognitive representations of performance paired with explicit evaluation of that performance may be doubly effective. This is especially true with evidence here and elsewhere that the outputs of monitoring processes are causally related to controlled behaviours.

5.3 Evidence Accumulation and Dual Process Theories

The present work is deeply rooted in each of evidence accumulation and dual-process accounts of reasoning and decision-making. Indeed, at times I link the two accounts to a degree that even the foremost proponents of each may not. For that reason, I'll briefly discuss how the two accounts may be reconciled. Modern accounts of dual-process reasoning (e.g., De Neys, 2023; Pennycook, Fugelsang, & Koehler, 2015) posit that individuals, over time, generate evidence in favour of two alternative choices. This position is shared with evidence accumulation accounts of two alternative forced choice tasks (see Evans & Wagenmakers, 2019 for a review). Modern accounts of dual-process reasoning also posit that one's confidence (or at least a sense of certainty or uncertainty), and thus controlled behaviour, arises from the relative difference in evidence at the time of a decision. This is also shared with current theorizing about evidence accumulation (e.g., Desender et al., 2021). In the present thesis, I have provided evidence that this metacognitive process is sensitive to variables that correspond conceptually with general evidence accumulation accounts. Specifically, items that are responded to faster on average and with a greater degree of response consensus are associated with, on average, higher confidence judgments. The account for these variables is that items whose cues align reach a decision threshold faster and with less competition from the alternative response. This is consistent with recent reasoning work.

One plausibly irreconcilable difference between dual-process models and EAMs is that at least some EAMs presuppose evidence accumulation is a stochastic process; cues are processed

essentially at random, or at least with large variability between individuals and/or across trials (Evans & Wagenmakers, 2019). Conversely, a theoretical tenet of dual-process theories is that some sources of evidence are processed faster or with the requirement of fewer resources. Refer to Figure 1, for example. The model presented involves the initial evaluation of options, and in conjunction with decades of dual-process theorizing, the intuitive option is advantaged early. However, a qualitatively distinct type of processing intervenes, at which point the analytic option is preferred. De Neys (2023) argues that modern dual-process theories are intractable; perhaps thinking about reasoning through an evidence accumulation lens can aid in establishing a path forward. The described pattern is argued to require the switch from one class of processing to another, and the resultant switch in favouring of one response to the other is an argued consequence. If such a pattern can be produced without the need for a qualitative shift in *processing* to produce an evidentiary shift between two responses, these accounts may not be best described by dual process frameworks. If, however, a qualitative shift in processing is necessary to produce such a pattern, it may be a path for dual-process theories to succeed. Either way, turning to a field where similar assumptions are made about individuals' choices may be a way to gain traction in the dual-process debate.

5.4 Limitations of Evidence and Generalizability

The present thesis suffers from a number of limitations, perhaps foremost of which is the nature of the evidence provided. Either outcome is theoretically interesting: either reasoning researchers have been systemically misestimating monitoring and control effects in reasoning by virtue of their methods of elicitation, or reasoning is a special case where method of elicitation is unimportant, unlike perceptual decision-making or memory. Given strong evidence in favour of either conclusion, a clear and interesting conclusion could have been made. The evidence I am

left with, however, is somewhat inconsistent. Variables that should be capturing similar constructs do not always correspond (although this is decidedly not unique to the present work); conditions in separate experiments that should be approximately equivalent are not. Though I have made my conclusions based on what I think the most reasonable interpretation of the data, these conclusions are made with less certainty than they otherwise might be. Perhaps future work probing the kind of questions I have approached here would be best suited first confirming the patterns I have identified.

Another possible limitation of the present work is the extent to which it is generalizable across tasks and across samples. All experiments contained in this thesis used a basic, syllogistic reasoning task. This task was particularly well-suited for the current purpose; the characteristics of this task are particularly amenable to the evaluation of metacognitive monitoring and its effects, and it is prominent in the literature for this reason. Also, there was an available set of items for which I had granular data describing one of the main characteristics by which to assess confidence. This allowed for the precise (or, at least, more precise than otherwise) modelling of the effects of believability on confidence. However, this modelling was done in a way that potentially limits is relevance beyond the current task and/or set of items. For example, all mixed-effects analyses were fit with only random intercepts nested within participants. A drawback of this kind of approach is that one is theoretically less able to draw broader conclusions outside of the present item set and sample. It may seem that this choice was an oversight or an error due to its limitation of generalizability, but it was actually made with generalizability in mind. A simpler model was chosen to decrease the odds that, when translated into an experimental setting, it would fail to converge or be otherwise mis-specified in one or both experimental conditions. That is, a random effects structure was chosen with full knowledge

that it is probably suboptimal, but because the focus of the present investigation was not optimally modelling monitoring behaviour. The chosen model structure allows for direct comparison of *a* model of confidence across comparable conditions, although it may not be the ideal model (the results of Chapter 3, however, suggest that this model is nevertheless pretty good). That said, should other researchers be interested in that endeavour, the conclusions presented herein may be useful. Further, the conclusions made in the present thesis, while mainly in the reasoning domain, are intended to be at least somewhat general. However, any degree of generality rests on the assumption that the processes observed in the present set of experiments translate to other tasks and domains. This assumption is not baseless; the field of reasoning, especially as it involves metacognitive processes, has a long history of adopting concepts and theories from other domains, largely with success.

5.5 Mechanistic Considerations and Future Directions

I observed monitoring and control differences as a result of differences in the method of measuring confidence, but nothing to suggest a reason or mechanism behind these effects. This work was inspired in large part by other work that, in different terms, has investigated internal versus external representations of confidence. For example, Shea and colleagues (2014) argue that there are dual metacognitive systems; one for the purpose of broadcasting monitoring outputs internally, and one for broadcasting them to others. In a similar vein, Bang and colleagues (2020) demonstrated that there are context-dependent mappings of metacognitive representations which facilitate social tuning of expressed metacognitive states. Either of these accounts (or both in conjunction) can provide an explanation of control differences; perhaps the outward registration of under-confidence prompts deliberation to a greater extent than does private evaluation of the same under-confidence. This would be consistent with control

differences found in Experiment 6. The observed monitoring effects, however, are more difficult to place in this context. Imagine, purely speculatively, that there are indeed separate metacognitive systems for broadcast internally versus externally. Might it be the case that the external broadcast system differentially weighs input signals based on what might be a useful signal to others? If that were true, a reasonable candidate to overweight might be one's subjective experience (i.e., the source of information to which one is solely privileged). This would be consistent with the monitoring effect observed in Experiment 6 where implicit measures of confidence were more sensitive to experienced ease or difficulty. Obviously, I have presented no evidence to this affect and as such these suggestions are purely exploratory. However, given metacognitive processes are operative in scenarios beyond simple syllogistic reasoning, such as social decision-making and cooperation, future work might benefit from consideration of the method of eliciting confidence.

5.6 Conclusions

Reasoning is ubiquitous in human life; from the earliest parts of your day, you are likely reasoning. The local meteorologist suggests it may take longer to get to work today; should you leave early? Your car is making an unusual noise; should you take it to the mechanic? Your uncle suggests you should cease chemotherapy and instead try colloidal silver; what treatment do you choose? Some instances of reasoning are (generally) inconsequential. Other instances have dramatic and long-lasting effects. Decades of psychological research suggest that humans process these types of situations by generating rapid intuitions, and only on some occasions supplanting those intuitions with more effortful thinking. But how do we decide which occasions to engage in deliberative thinking? Recent work suggests that this process may be a targeted process of monitoring one's intuitions and engaging in top-down, directed processing pending

the outcome of that monitoring process. Critical, then, is the harnessing of this process to avoid pitfalls in – at least – the most crucial decisions.

The work presented here suggests that explicit evaluation of one's intuition not only alters the basis of deliberative thinking, but also the basis on which a sense of certainty is formed. Particularly uncertain explicit evaluations prompt more effortful processing than organic feelings of uncertainty. In addition, explicit evaluations seem to alter the basis of evaluation in the first place, undervaluing heuristic, undiagnostic, ease-based factors (silver just *sounds* better than Cyclophosphamide, after all). The work presented here is not suggestive of judgment reactivity as an intervention of any sort, but it does suggest some degree of previously unidentified reactivity in human reasoning. At a minimum, there are underappreciated methodological considerations in current investigations of human reasoning. Future work can establish the range and severity of these. More suggestively, my work indicates that the explicit evaluation of confidence causes predictable changes in the reasoning process that may be of widespread significance. The current work represents an important step toward the precise characterization of monitoring and control processes in reasoning.

References

Ackerman, R., & Thompson, V. A. (2017). Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617. https://doi.org/10.1016/j.tics.2017.05.004

Adam, K. C., & Vogel, E. K. (2016). Reducing failures of working memory with performance feedback. *Psychonomic Bulletin & Review*, *23*, 1520-1527. https://doi.org/10.3758/s13423-016-1019-4

Alexander, W. H., & Brown, J. W. (2010). Computational models of performance monitoring and cognitive control. *Topics in cognitive science*, *2*(4), 658-677. https://doi.org/10.1111/j.1756-8765.2010.01085.x

Alter, A. L. (2013). The benefits of cognitive disfluency. *Current Directions in Psychological Science*, *22*(6), 437-442. https://doi.org/10.1177/09637214134988

Alter, A. L., Oppenheimer, D. M., & Epley, N. (2013). Disfluency prompts analytic thinking—But not always greater accuracy: Response to. *Cognition*, *128*(2), 252-255. https://doi.org/10.1016/j.cognition.2013.01.006

Arnold, M. M., & Prike, T. (2015). Comparative difficulty and the strategic regulation of accuracy: The impact of test-list context on monitoring and meta-metacognition. *Acta Psychologica*, *157*, 155-163. https://doi.org/10.1016/j.actpsy.2015.02.018

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

Bajšanski, I., Žauhar, V., & Valerjev, P. (2019). Confidence judgments in syllogistic reasoning: The role of consistency and response cardinality. *Thinking & Reasoning*, *25*(1), 14-47. https://doi.org/10.1080/13546783.2018.1464506

Bang, D., Ershadmanesh, S., Nili, H., & Fleming, S. M. (2020). Private–public mappings in human prefrontal cortex. *Elife*, *9*, e56477. https://doi.org/10.7554/eLife.56477

Barrett T, Dowle M, Srinivasan A, Gorecki J, Chirico M, Hocking T (2024). _data.table: Extension of `data.frame`_. R package version 1.15.4, <https://CRAN.R-project.org/package=data.table>.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Batha, K., & Carroll, M. (2007). Metacognitive training aids decision making. *Australian Journal of Psychology*, *59*(2), 64-69. https://doi.org/10.1080/00049530601148371

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & cognition*, *31*, 297-305. https://doi.org/10.3758/BF03194388

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55. https://doi.org/10.1037/0096-3445.127.1.55

Beran, M. J., Parrish, A. E., Perdue, B. M., & Washburn, D. A. (2014). Comparative cognition: Past, present, and future. *International journal of comparative psychology*, *27*(1), 3. https://doi.org/10.46867/ijcp.2014.27.01.07

Beran, M. J., Perdue, B. M., Futch, S. E., Smith, J. D., Evans, T. A., & Parrish, A. E. (2015). Go when you know: Chimpanzees' confidence movements reflect their responses in a computerized memory task. *Cognition*, *142*, 236-246. https://doi.org/10.1016/j.cognition.2015.05.023

Besken, M., & Mulligan, N. W. (2013). Easily perceived, easily remembered? Perceptual interference produces a double dissociation between metamemory and memory performance. *Memory & cognition*, *41*, 897-903. https://doi.org/10.3758/s13421-013-0307-8

Besken, M., & Mulligan, N. W. (2014). Perceptual fluency, auditory generation, and metamemory: analyzing the perceptual fluency hypothesis in the auditory modality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 429. https://doi.org/10.1037/a0034407

Białek, M., Muda, R., Stewart, K., Niszczota, P., & Pieńkosz, D. (2020). Thinking in a foreign language distorts allocation of cognitive effort: Evidence from reasoning. *Cognition*, *205*, 104420. https://doi.org/10.1016/j.cognition.2020.104420

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, *64*(1), 417-444. https://doi.org/10.1146/annurev-psych-113011-143823

Boldt, A., & Gilbert, S. J. (2022). Partially overlapping neural correlates of metacognitive monitoring and metacognitive control. *Journal of Neuroscience*, *42*(17), 3622-3635. https://doi.org/10.1523/JNEUROSCI.1326-21.2022

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological review*, *108*(3), 624. https://doi.org/10.1037/0033-295X.108.3.624

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. (2023). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*, *7*(1), 142-154. https://doi.org/10.1038/s41562-022-01464-x

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, *10*(4), 433-436.

Brewer, W. F., & Sampaio, C. (2006). Processes leading to confidence and accuracy in sentence recognition: A metamemory approach. *Memory*, *14*(5), 540-552. https://doi.org/10.1080/09658210600590302

Brewer, W. F., & Sampaio, C. (2012). The metamemory approach to confidence: A test using semantic memory. *Journal of Memory and Language*, *67*(1), 59-77. https://doi.org/10.1016/j.jml.2012.04.002

Brisson, J., Schaeken, W., Markovits, H., & De Neys, W. (2018). Conflict detection and logical complexity. *Psychologica Belgica*, *58*(1), 318. https://doi.org/10.5334/pb.448

Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. *Acta psychologica*, *197*, 153-165. https://doi.org/10.1016/j.actpsy.2019.04.011

Buratti, S., Allwood, C. M., & Kleitman, S. (2013). First-and second-order metacognitive judgments of semantic memory reports: The influence of personality traits and cognitive styles. *Metacognition and learning*, *8*, 79-102. https://doi.org/10.1007/s11409-013-9096-5

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review*, *100*(3), 432. https://doi.org/10.1037/0033-295X.100.3.432

Carruthers, P., & Williams, D. M. (2019). Comparative metacognition. *Animal Behavior and Cognition, 6*(4)*,* 278-288. https://doi.org/10.26451/abc.06.04.08.2019

Carruthers, P., & Williams, D. M. (2022). Model-free metacognition. *Cognition*, *225*, 105-117. https://doi.org/10.1016/j.cognition.2022.105117

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & cognition*, *36*, 429-437. https://doi.org/10.3758/MC.36.2.429

Castel, A.D., McCabe, D.P. & Roediger, H.L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review, 14*, 107–111. https://doi.org/10.3758/BF03194036

Caziot, B., & Mamassian, P. (2021). Perceptual confidence judgments reflect self-consistency. *Journal of Vision*, *21*(12), 8-8. https://doi.org/10.1167/jov.21.12.8

Cowan, N., Hardman, K., Saults, J. S., Blume, C. L., Clark, K. M., & Sunday, M. A. (2016). Detection of the number of changes in a display in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(2), 169–185. https://doi.org/10.1037/xlm0000163

de Bruin, A. B., Kok, E. M., Lobbestael, J., & de Grip, A. (2017). The impact of an online tool for monitoring and regulating learning at university: Overconfidence, learning strategy, and personality. *Metacognition and Learning*, *12*, 21-43. https://doi.org/10.1007/s11409-016-9159-5

De Neys, W. (2006a). Dual processing in reasoning: Two systems but one reasoner. *Psychological science*, *17*(5), 428-433. https://doi.org/10.1111/j.1467-9280.2006.01723.x

De Neys, W. (2006b). Automatic–heuristic and executive–analytic processing during reasoning: Chronometric and dual-task considerations. *Quarterly journal of experimental psychology*, *59*(6), 1070-1100. https://doi.org/10.1080/027249805430001

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). Bias, conflict, and fast logic: Towards a hybrid dual process future? In *Dual process theory 2.0* (pp. 47-65). Routledge.

De Neys, W. (2023). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, *46*, e111. https://doi.org/10.1017/S0140525X2200142X

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

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological science*, *29*(5), 761-778. https://doi.org/10.1177/0956797617744771

Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition*, *207*, 104522. https://doi.org/10.1016/j.cognition.2020.104522

Double, K. S., & Birney, D. P. (2019). Reactivity to measures of metacognition. *Frontiers in Psychology*, *10*, 2755. https://doi.org/10.3389/fpsyg.2019.02755

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory*, *26*(6), 741-750. https://doi.org/10.3389/fpsyg.2019.02755

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & cognition*, *33*, 1096-1115. https://doi.org/10.3758/BF03193216

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Sage Publications.

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

Evans, J. St. B. T. (2010). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry*, *21*(4), 313-326. https://doi.org/10.1080/1047840X.2010.521057

Evans, J. St. B. T. (2019). Reflections on reflection: the nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, *25*(4), 383-415. https://doi.org/10.1080/13546783.2019.1623071

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, *11*(3), 295-306. https://doi.org/10.3758/BF03196976

Evans, J. St. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, *11*(4), 382-389. https://doi.org/10.1080/13546780542000005

Evans, J. St. B. T., Handley, S. J., & Harper, C. N. (2001). Necessity, possibility and belief: A study of syllogistic reasoning. *The Quarterly Journal of Experimental Psychology: Section A*, *54*(3), 935-958. https://doi.org/10.1080/713755983

Evans, J. St. 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

Evans, J. St. B. T., & Wason, P. C. (1976). Rationalization in a reasoning task. *British Journal of Psychology*, *67*(4), 479-486. https://doi.org/10.1111/j.2044-8295.1976.tb01536.x

Evans, N. J., & Wagenmakers, E. J. (2019). Evidence accumulation models: Current limitations and future directions. PsyArXiv, Published online May 16, 2019. https://doi.org/10.31234/osf.io/74df9

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191. https://doi.org/10.3758/BF03193146

Fiacconi, C. M., & Dollois, M. A. (2020). Does word frequency influence judgments of learning (JOLs)? A meta-analytic review. *Canadian Journal of Experimental Psychology, 74*(4), 346–353. https://doi.org/10.1037/cep0000206

Fiacconi, C. M., Mitton, E. E., Laursen, S. J., & Skinner, J. (2020). Isolating the contribution of perceptual fluency to judgments of learning (JOLs): Evidence for reactivity in measuring the influence of fluency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*(5), 926–944. https://doi.org/10.1037/xlm0000766

Finley, J. R., Tullis, J. G., & Benjamin, A. S. (2010). Metacognitive control of learning and remembering. *New science of learning: Cognition, computers and collaboration in education*, 109-131. https://doi.org/10.1007/978-1-4419-5716-0_6

Fitzgerald, L. M., Arvaneh, M., & Dockree, P. M. (2017). Domain-specific and domain-general processes underlying metacognitive judgments. *Consciousness and Cognition*, *49*, 264-277. https://doi.org/10.1016/j.concog.2017.01.011

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, *75*(1), 241-268. https://doi.org/10.1146/annurev-psych-022423-032425

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*(1), 91–114. https://doi.org/10.1037/rev0000045

Fleming, S. M., & Dolan, R. J. (2010). Effects of loss aversion on post-decision wagering: Implications for measures of awareness. *Consciousness and cognition*, *19*(1), 352-363. https://doi.org/10.1016/j.concog.2009.11.002

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical transactions of the royal society B: Biological sciences*, *367*(1594), 1280-1286. https://doi.org/10.1098/rstb.2012.0021

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *Journal of Neuroscience*, *32*(18), 6117-6125. https://doi.org/10.1523/JNEUROSCI.6489-11.2012

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in human neuroscience*, *8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, *137*(10), 2811-2822. https://doi.org/10.1093/brain/awu221

Frankish, K., & Evans, J. St. B. T. (2009). The duality of mind: An historical perspective. *In two minds: Dual processes and beyond*, 1-29.

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning*, *15*(2), 105–128. https://doi.org/10.1080/13546780802711185

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

Ghasemi, O., Handley, S., Howarth, S., Newman, I. R., & Thompson, V. A. (2022). Logical intuition is not really about logic. *Journal of Experimental Psychology: General*, *151*(9), 2009-2028. https://doi.org/10.31219/osf.io/gxd73

Giner-Sorolla, R., Montoya, A. K., Reifman, A., Carpenter, T., Lewis Jr, N. A., Aberson, C. L., ... & Soderberg, C. (2024). Power to detect what? Considerations for planning and evaluating sample size. *Personality and Social Psychology Review*, *28*(3), 276-301. https://doi.org/10.1177/10888683241228328

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms?. *Comparative cognition & behavior reviews*, *4*, 17. https://doi.org/10.3819/ccbr.2009.40002

Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. In *Psychology of learning and motivation* (Vol. 62, pp. 33-58). Academic Press. https://doi.org/10.1016/bs.plm.2014.09.002

Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of risk and uncertainty*, *4*(1), 5-28. https://doi.org/10.1007/BF00057884

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, memory, and cognition*, *34*(5), 1191. https://doi.org/10.1037/a0013025

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(1), 22–34. https://doi.org/10.1037/0278-7393.29.1.22

Hines, J. C., Touron, D., & Hertzog, C. (2009). Metacognitive influences on study time allocation in an associative recognition task: An analysis of adult age differences. *Psychology & Aging, 24*, 462-475. doi:10.1037/a0014417

Holroyd, C. B., Yeung, N., Coles, M. G., & Cohen, J. D. (2005). A mechanism for error detection in speeded response time tasks. *Journal of Experimental Psychology: General*, *134*(2), 163. https://doi.org/10.1037/0096-3445.134.2.163

Insabato, A., Pannunzi, M., Rolls, E. T., & Deco, G. (2010). Confidence-related decision making. *Journal of neurophysiology*, *104*(1), 539-547. https://doi.org/10.1152/jn.01068.2009

Janes, J. L., Rivers, M. L., & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: Positive, negative, or both?. *Psychonomic bulletin & review*, *25*, 2356-2364. https://doi.org/10.3758/s13423-018-1463-4

Jang, Y., & Nelson, T. O. (2005). How Many Dimensions Underlie Judgments of Learning and Recall? Evidence From State-Trace Methodology. *Journal of Experimental Psychology: General, 134*(3), 308–326. https://doi.org/10.1037/0096-3445.134.3.308

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

Kelemen, W. L., & Weaver III, C. A. (1997). Enhanced memory at delays: Why do judgments of learning improve over time?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(6), 1394. https://doi.org/10.1037/0278-7393.23.6.1394

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1322-1337. https://doi.org/10.1098/rstb.2012.0037

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on psychological science*, *4*(6), 533-550. https://doi.org/10.1111/j.1745-6924.2009.01164.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759-764. https://doi.org/10.1126/science.1169405

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., Broussard, C., et al. (2007). What's new in Psychtoolbox-3. *Perception, 36*, 1.

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and individual differences*, *17*(2), 161-173. https://doi.org/10.1016/j.lindif.2007.03.004

Konstantinidis, E., & Shanks, D. R. (2014). Don't bet on it! Wagering as a measure of awareness in decision making under uncertainty. *Journal of Experimental Psychology: General*, *143*(6), 2111. https://doi.org/10.1037/a0037977

Koriat, A. (1976). Another look at the relationship between phonetic symbolism and the feeling of knowing. *Memory & Cognition*, *4*(3), 244-248. https://doi.org/10.3758/BF03213170

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological review*, *100*(4), 609. https://doi.org/10.1037/0033-295X.100.4.609

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289–325). Cambridge University Press. https://doi.org/10.1017/CBO9780511816789.012

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(4), 945–959. https://doi.org/10.1037/0278-7393.34.4.945

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological review*, *119*(1), 80. https://doi.org/10.1037/a0025648

Koriat, A., & Bjork, R. A. (2005). Illusions of Competence in Monitoring One's Knowledge During Study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 187–194. https://doi.org/10.1037/0278-7393.31.2.187

Koriat, A., Bjork, R.A. (2006) Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition, 34*, 959–972. https://doi.org/10.3758/BF03193244

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting One's Own Forgetting: The Role of Experience-Based and Theory-Based Processes.*Journal of Experimental Psychology: General, 133*(4), 643–656. https://doi.org/10.1037/0096-3445.133.4.643

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. Psychological Review, 103(3), 490–517. https://doi.org/10.1037/0033-295X. 103.3.490

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of memory and Language*, *52*(4), 478-492. https://doi.org/10.1016/j.jml.2005.01.001

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning.*Journal of Experimental Psychology: General, 138*(4), 449–468. https://doi.org/10.1037/a0017350

Kosourikhina, V. (2024). Uncertainty monitoring in reasoning: cue consistency is more important than belief-logic conflict. PsyArXiv, Published online November 18, 2024. https://doi.org/10.31234/osf.io/nds9h

Kumle, L., Võ, M. L. H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior research methods*, *53*(6), 2528-2543. https://doi.org/10.3758/s13428-021-01546-0

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lau, H. (2019) Consciousness, metacognition, & perceptual reality monitoring. PsyArXiv, Published online June 10, 2019. https://dx.doi.org/10.31234/osf.io/ckbyf

Lawrence MA (2016). _ez: Easy Analysis and Visualization of Factorial Experiments_. R package version 4.4-0, <https://CRAN.R-project.org/package=ez>.

Littrell, S., Fugelsang, J., & Risko, E. F. (2020). Overconfidently underthinking: Narcissism negatively predicts cognitive reflection. *Thinking & Reasoning*, *26*(3), 352-380. https://doi.org/10.1080/13546783.2019.1633404

Lobato, E., Mendoza, J., Sims, V., & Chin, M. (2014). Examining the relationship between conspiracy theories, paranormal beliefs, and pseudoscience acceptance among a university population. *Applied Cognitive Psychology*, *28*(5), 617-625. https://doi.org/10.1002/acp.3042

Lüdecke D (2024). _sjPlot: Data Visualization for Statistics in Social Science_. R package version 2.8.16, <https://CRAN.R-project.org/package=sjPlot>.

Markovits, H., Thompson, V. A., & Brisson, J. (2015). Metacognition and abstract reasoning. *Memory & cognition*, *43*, 681-693. https://doi.org/10.3758/s13421-014-0488-9

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs?. *Memory & Cognition*, *29*, 222-233. https://doi.org/10.3758/BF03194916

McCabe, D. P., & Soderstrom, N. C. (2011). Recollection-based prospective metamemory judgments are more accurate than those based on confidence: Judgments of remembering and knowing (JORKS). *Journal of Experimental Psychology: General, 140*(4), 605–621. https://doi.org/10.1037/a0024014

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., De Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *Journal of Neuroscience*, *33*(5), 1897-1906. https://doi.org/10.1523/JNEUROSCI.1890-12.2013

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in cognitive sciences*, *22*(4), 280-293. https://doi.org/10.1016/j.tics.2018.02.001

Metcalfe, J. (1986). Feeling of knowing in memory and problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(2), 288–294. https://doi.org/10.1037/0278-7393.12.2.288

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current directions in psychological science*, *18*(3), 159-163. https://doi.org/10.1111/j.1467-8721.2009.01628

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174-179. https://doi.org/10.3758/PBR.15.1.174

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of memory and language*, *52*(4), 463-477. https://doi.org/10.1016/j.jml.2004.12.001

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., ... & Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General*, *144*(2), e16. https://doi.org/10.1037/xge0000049

Meyers, E. A., Gretton, J. D., Budge, J. R., Fugelsang, J. A., & Koehler, D. J. (2023). Broad effects of shallow understanding: Explaining an unrelated phenomenon exposes the illusion of explanatory depth. *Judgment and Decision Making*, *18*, e24. https://doi.org/10.1017/jdm.2023.24

Miele, D. B., Finn, B., & Molden, D. C. (2011). Does easily learned mean easily remembered? It depends on your beliefs about intelligence. *Psychological Science*, *22*(3), 320-324. https://doi.org/10.1177/09567976103979

Miele, D. B., Son, L. K., & Metcalfe, J. (2013). Children's naive theories of intelligence influence their metacognitive judgments. *Child development*, *84*(6), 1879-1886. https://doi.org/10.1111/cdev.12101

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*(2), 200–219. https://doi.org/10.1037/a0039923

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological methods*, *16*(4), 406. https://doi.org/10.1037/a0024377

Morey R, Rouder J (2024). _BayesFactor: Computation of Bayes Factors for Common Designs_. R package version 0.9.12-4.7, <https://CRAN.R-project.org/package=BayesFactor>.

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2016). The effect of identical word pairs on people's metamemory judgments: What are the contributions of processing fluency and beliefs about memory? *Quarterly Journal of Experimental Psychology*, *69*(4), 781-799. https://doi.org/10.1080/17470218.2015.1058404

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory?. *Journal of Memory and Language*, *70*, 1-12. https://doi.org/10.1016/j.jml.2013.09.007

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic bulletin & review*, *20*, 378-384. https://doi.org/10.3758/s13423-012-0343-6

Murphy, D. H., Huckins, S. C., Rhodes, M. G., & Castel, A. D. (2022). The effect of perceptual processing fluency and value on metacognition and remembering. *Psychonomic Bulletin & Review*, *29*(3), 910-921. https://doi.org/10.3758/s13423-021-02030-8

Myers, S. J. (2023). *Is Judgment Reactivity Really about the Judgment?* (Publication No. 30490919) [Doctoral dissertation, Colorado State University]. ProQuest Dissertations & Theses.

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & Cognition*, *48*, 745-758. https://doi.org/10.3758/s13421-020-01025-5

Navarro, D. J. (2015) Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.6) University of New South Wales. Sydney, Australia

Nelson, T. O. (1996). Consciousness and metacognition. *American psychologist*, *51*(2), 102.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*(4), 267-271.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition. *Metacognition: Knowing About Knowing, 13*, 1–25. https://doi.org/10.7551/mitpress/4561.003.0003

Overgaard, M., & Sandberg, K. (2012). Kinds of access: different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1287-1296. https://doi.org/10.1098/rstb.2011.0425

Pennycook, G. (2023). A framework for understanding reasoning errors: From fake news to climate change and beyond. In *Advances in experimental social psychology* (Vol. 67, pp. 131-208). Academic Press. https://doi.org/10.1016/bs.aesp.2022.11.003

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision making*, *10*(6), 549-563. https://doi.org/10.1017/S1930297500006999

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., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39-50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 544. https://doi.org/10.1037/a0034887

Perrin, D., Sant'Anna, A. (2022). Episodic memory and the feeling of pastness: from intentionalism to metacognition. *Synthese 200*, 109. https://doi.org/10.1007/s11229-022-03567-4

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature neuroscience*, *10*(2), 257-261. https://doi.org/10.1038/nn1840

Pescetelli, N., Rees, G., & Bahrami, B. (2016). The perceptual and social components of metacognition.*Journal of Experimental Psychology: General, 145*(8), 949–965. https://doi.org/10.1037/xge0000180

Proust, J. (2012). Metacognition and mindreading: one or two functions. *Foundations of metacognition*, 234-251.

Prowse Turner, J. A., & Thompson, V. A. (2009). The role of training, alternative models, and logical necessity in determining confidence in syllogistic reasoning. *Thinking & Reasoning*, *15*(1), 69–100. https://doi.org/10.1080/13546780802619248

R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Rademaker, R. L., Tredway, C. H., & Tong, F. (2012). Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *Journal of vision*, *12*(13), 21-21. https://doi.org/10.1167/12.13.21

Rahnev, D. (2021). Visual metacognition: Measures, models, and neural correlates. *American Psychologist, 76*(9), 1445–1453. https://doi.org/10.1037/amp0000937

Raoelison, M., Boissin, E., Borst, G., & De Neys, W. (2021). From slow to fast logic: The development of logical intuitions. *Thinking & Reasoning*, *27*(4), 599-622. https://doi.org/10.1080/13546783.2021.1885488

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in cognitive sciences*, *20*(4), 260-281. https://doi.org/10.1016/J.TICS.2016.01.007

Rawson, K. A., O'Neil, R., & Dunlosky, J. (2011). Accurate monitoring leads to effective control and greater learning of patient education materials. *Journal of Experimental Psychology: Applied*, *17*(3), 288. https://doi.org/10.1037/a0024749

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience?. *Personality and social psychology review*, *8*(4), 364-382. https://doi.org/10.1207/s15327957pspr0804_3

Revelle, W. (2024). _psych: Procedures for Psychological, Psychometric, and Personality Research_. Northwestern University, Evanston, Illinois. R package version 2.4.6, <https://CRAN.R-project.org/package=psych>.

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. *The Oxford handbook of metamemory*, *1*, 65-80.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*(4), 615–625. https://doi.org/10.1037/a0013684

Rivers, M. L., Janes, J. L., & Dunlosky, J. (2021). Investigating memory reactivity with a within-participant manipulation of judgments of learning: Support for the cue-strengthening hypothesis. *Memory*, *29*(10), 1342-1353. https://doi.org/10.1080/09658211.2021.1985143

Rouault M, McWilliams A, Allen M. G., Fleming S. M. (2018) Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personality Neuroscience*. 2018;1:e17. doi:10.1017/pen.2018.16

Sá, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 91*(3), 497–510. https://doi.org/10.1037/0022-0663.91.3.497

Schacter, D. L. (1983). Feeling of knowing in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(1), 39–54 https://doi.org/10.1037/0278-7393.9.1.39

Schneider, S. L., & Lopes, L. L. (1986). Reflection in preferences under risk: Who and when may suggest why. *Journal of Experimental Psychology: Human perception and performance*, *12*(4), 535. https://doi.org/10.1037/0096-1523.12.4.535

Schurger, A., & Sher, S. (2008). Awareness, loss aversion, and post-decision wagering. *Trends in Cognitive Sciences*, *12*, 209–210. https://doi.org/10.1016/j.tics.2008.02.012

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193. https://doi.org/10.1016/j.tics.2014.01.006

Shekhar, M., & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review, 128*(1), 45-70 https://doi.org/10.1037/rev0000249

Shin, J., & Ariely, D. (2004). Keeping doors open: The effect of unavailability on incentives to keep options viable. *Management science*, *50*(5), 575-586. https://doi.org/10.1287/mnsc.1030.0148

Shynkaruk, J.M., Thompson, V.A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition, 34*, 619–632. https://doi.org/10.3758/BF03193584

Sidi, Y., Ackerman, R. (2024). Opting Out as an Untapped Resource in Instructional Design: Review and Implications. *Educational Psychology Review*, *36*, 41. https://doi.org/10.1007/s10648-024-09879-w

Sidi, Y., Shpigelman, M., Zalmanov, H., & Ackerman, R. (2017). Understanding metacognitive inferiority on screen by exposing cues for depth of processing. *Learning and instruction*, *51*, 61-73. https://doi.org/10.1016/j.learninstruc.2017.01.002

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

Smith, J. D., Couchman, J. J., & Beran, M. J. (2014). Animal metacognition: a tale of two comparative psychologies. *Journal of Comparative Psychology*, *128*(2), 115. https://doi.org/10.1037/a0033105

Smith, J. D., Beran, M. J., Redford, J. S., & Washburn, D. A. (2006). Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. *Journal of Experimental Psychology: General, 135*(2), 282–297 https://doi.org/10.1037/0096-3445.135.2.282

Smith, T. R., Parrish, A. E., Creamer, C., Rossettie, M., & Beran, M. J. (2020). Capuchin monkeys (sometimes) go when they know: Confidence movements in Sapajus apella. *Cognition*, *199*, 104237. https://doi.org/10.1016/j.cognition.2020.104237

Snodgrass, M., Kalaida, N., & Winer, E. S. (2009). Access is mainly a second-order process: SDT models whether phenomenally (first-order) conscious states are accessed by reflectively (second-order) conscious processes. *Consciousness and Cognition*, *18*(2), 561-564. https://doi.org/10.1016/j.concog.2009.01.003

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 553. https://doi.org/10.1037/a0038388

Solcz, S. (2023). *Not All Syllogisms Are Created Equal: Varying Premise Believability Reveals Differences Between Conditional and Categorical Syllogisms.* (Publication No. 973346542) [Doctoral dissertation, University of Waterloo]. Theses Canada.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204. https://doi.org/10.1037/0278-7393.26.1.204

Song, J. H., Loyal, S., & Lond, B. (2021). Metacognitive awareness scale, domain specific (MCAS-DS): Assessing metacognitive awareness during Raven's progressive matrices. *Frontiers in Psychology*, *11*, 607577. https://doi.org/10.3389/fpsyg.2020.607577

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. *The Oxford handbook of thinking and reasoning*, 343-365. https://doi.org/10.1093/oxfordhb/9780199734689.013.0022

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of personality and social psychology*, *94*(4), 672. https://doi.org/10.1037/0022-3514.94.4.672

Steinhauser, M., Maier, M., & Hübner, R. (2008). Modeling behavioral measures of error detection in choice tasks: response monitoring versus conflict monitoring. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(1), 158. https://doi.org/10.1037/0096-1523.34.1.158

Stewart, K. M., Risko, E. F., & Fugelsang, J. (2023). Response generation, not response execution, influences feelings of rightness in reasoning. *Quarterly Journal of Experimental Psychology*, *76*(10), 2379-2389. https://doi.org/10.1177/17470218231156712

Susser, J. A., Mulligan, N. W., & Besken, M. (2013). The effects of list composition and perceptual fluency on judgments of learning (JOLs). *Memory & Cognition*, *41*, 1000-1011. https://doi.org/10.3758/s13421-013-0323-8

Tauber, S. K., & Rhodes, M. G. (2012). Measuring memory monitoring with judgements of retention (JORs). *Quarterly Journal of Experimental Psychology, 65*(7), 1376-1396. https://doi.org/10.1080/17470218.2012.656665

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 1024–1037. https://doi.org/10.1037/0278-7393.25.4.1024

Thompson, V. A., Ackerman, R., Sidi, Y., Ball, L. J., Pennycook, G., & Turner, J. A. P. (2013). The role of answer fluency and perceptual fluency in the monitoring and control of reasoning: Reply to. *Cognition*, *128*(2), 256-258. https://doi.org/10.1016/j.cognition.2013.03.003

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

Tiedemann F (2022). _gghalves: Compose Half-Half Plots Using Your Favourite Geoms_. R package version 0.1.4, <https://CRAN.R-project.org/package=gghalves>.

Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: consciousness as an unconscious re-description process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1412-1423. https://doi.org/10.1098/rstb.2011.0421

Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & cognition*, *45*, 539-552. https://doi.org/10.3758/s13421-016-0680-1

Tullis, J. G., & Benjamin, A. S. (2012). Consequences of restudy choices in younger and older learners. *Psychonomic Bulletin & Review*, *19*, 743-749. https://doi.org/10.3758/s13423-012-0266-2

Undorf, M., Livneh, I., & Ackerman, R. (2021). Metacognitive control processes in question answering: Help seeking and withholding answers. *Metacognition and Learning*, *16*(2), 431-458. https://doi.org/10.1007/s11409-021-09259-7

Undorf, M., Schäfer, F., & Halamish, V. (2024). Making Judgments of Learning Either Enhances or Impairs Memory: Evidence From 17 Experiments With Related and Unrelated Word Pairs. *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.117108

Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & cognition*, *46*, 507-519. https://doi.org/10.3758/s13421-017-0780-6

Undorf, M., Zimdahl, M. F., & Bernstein, D. M. (2017). Perceptual fluency contributes to effects of stimulus size on judgments of learning. *Journal of Memory and Language*, *92*, 293-304. https://doi.org/10.1016/j.jml.2016.07.003

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550. https://doi.org/10.1037/0033-295X.108.3.550

van Langen, J. (2020). Open-visualizations in R and Python. https://github.com/jorvlan/openvisualizations

Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1147. https://doi.org/10.1037/0278-7393.30.6.1147

Wang, S., & Thompson, V. (2019). Fluency and feeling of rightness: The effect of anchoring and models. *Psihologijske teme*, *28*(1), 37-72. https://doi.org/10.31820/pt.28.1.3

Watson, C. S., Kellogg, S. C., Kawanishi, D. T., & Lucas, P. A. (1973). The uncertain response in detection-oriented psychophysics. *Journal of Experimental Psychology, 99*(2), 180–185 https://doi.org/10.1037/h0034736

Whatley, M.C., Castel, A.D. (2022). The role of metacognition and schematic support in younger and older adults' episodic memory. *Memory & Cognition, 50*, 601–616, https://doi.org/10.3758/s13421-021-01169-y

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of statistical software*, *40*, 1-29. https://doi.org/10.18637/jss.v040.i01

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of open source software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Witherby, A. E., & Tauber, S. K. (2017). The concreteness effect on judgments of learning: Evaluating the contributions of fluency and beliefs. *Memory & cognition*, *45*, 639-650. https://doi.org/10.3758/s13421-016-0681-0

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310-1321. https://doi.org/10.1098/rstb.2011.0416

Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of confidence. *Journal of Experimental Psychology: General*, *144*(2), 489. https://doi.org/10.1037/xge0000062

Appendix A: Items Used in All Experiments

Categorical Syllogisms

Syllogism	Validity	Average Premise Believability (/5)	Conclusion Believability (/5)
No actresses are men Some singers are men Therefore some singers are not actresses	Valid	4.545	4.44
All slacks are pants All pants are clothing Therefore all slacks are clothing	Valid	4.115	3.92
No dogs are felines Some furry animals are felines Therefore some furry animals are not dogs	Valid	3.89	4.50
No birds are salmon Some fish are salmon Therefore some fish are not birds	Valid	4.555	2.67
All rats are pests All pests are insects Therefore All rats are insects	Valid	2.095	1.00
No puppies are cobras Some snakes are cobras Therefore some snakes are not puppies	Valid	4.585	2.87
No black animals are rodents Some cats are rodents Therefore some cats are not black animals	Valid	1.69	4.02
No poisonous animals are insects Some snakes are insects Therefore some snakes are not poisonous	Valid	1.475	4.76
All reptiles are rabbits All rabbits are cold-blooded Therefore all reptiles are coldblooded	Valid	1.56	4.19

Syllogism	Validity	Average Premise Believability (/5)	Conclusion Believability (/5)
All cherries are tools All tools are wooden things Therefore all cherries are wooden things	Valid	1.59	1.21
No animals are dogs Some lions are dogs Therefore some lions are not animals	Valid	1.23	1.16
No flowers are daisies Some daffodils are daisies Therefore some daffodils are not flowers	Valid	1.805	1.81
No grandparents are children Some females are children Therefore some grandparents are not females	Invalid	4.345	4.54
All infants are newborns All newborns are babies Therefore all babies are infants	Invalid	3.88	3.79
No teenagers are elderly Some males are elderly Therefore some teenagers are not males	Invalid	4.46	4.52
No roses are weeds Some plants are weeds Therefore some roses are not plants	Invalid	4.055	1.94
All toddlers are children All children are humans Therefore all humans are toddlers	Invalid	4.205	1.79
No cod are angelfish Some fish are angelfish Therefore some cod are not fish	Invalid	4.34	1.43
No books are novels Some magazines are novels Therefore some books are not magazines	Invalid	1.89	3.98

Syllogism	Validity	Average Premise Believability (/5)	Conclusion Believability (/5)
All copper things are toys All toys are pennies Therefore all pennies are copper things	Invalid	1.56	3.99
No vehicles are buses Some convertibles are buses Therefore some vehicles are not convertibles	Invalid	1.685	4.86
No monkeys are primates Some alligators are primates Therefore some alligators are not monkeys	Invalid	1.405	2.31
No butterflies are insects Some rabbits are insects Therefore some butterflies are not rabbits	Invalid	1.415	2.69
All diamonds are inexpensive All inexpensive things are gemstones Therefore all gemstones are diamonds	Invalid	1.715	1.80

Conditional Syllogisms

Syllogism	Validity	Average Premise Believability (/5)	Conclusion Believability (/5)
If a room is dirty then it is not clean If a room is not clean then it is messy Therefore if a room is dirty then it is messy	Valid	4.055	3.89
If an animal is a tiger then it eats meat If an animal eats meat then it is a carnivore Therefore if an animal is a tiger then it is a carnivore	Valid	3.95	4.23
If an animal is a bird then it has a beak If an animal has a beak then it has feathers Therefore if an animal is a bird then it has feathers	Valid	4.09	3.85
If an animal is a whale then it is a mammal If an animal is a mammal then it is a land-dweller Therefore if an animal is a whale then it is a land-dweller	Valid	3.915	1.34
If a person is an employee then she is not retired If a person is not retired then she is young Therefore if a person is an employee then she is young	Valid	3.78	1.50
If a vehicle is a motorcycle then it has wheels If a vehicle has wheels then it has a steering wheel Therefore if a vehicle is a motorcycle then it has a steering wheel	Valid	3.98	2.13
If a food is a chilli pepper then it is not a vegetable If a food is not a vegetable then it is spicy Therefore if a food is a chilli pepper then it is spicy	Valid	1.90	4.19
If a person is a pianist then he is a professor If a person is a professor then he is a musician Therefore if a person is a pianist then he is a musician	Valid	1.935	4.21
If an animal is a puppy then it is cold-blooded If an animal is cold-blooded then it is a canine Therefore if an animal is a puppy then it is a canine	Valid	1.365	3.92

Syllogism	Validity	Average Premise Believability (/5)	Conclusion Believability (/5)
If an animal is a flamingo then it is a cat If an animal is a cat then it is a reptile Therefore if an animal is a flamingo then it is a reptile	Valid	1.65	1.48
If a person is a father then they are a doctor If a person is a doctor then they are a woman Therefore if a person is a father then they are a woman	Valid	3.10	1.34
If an object is a musical instrument then it is not a piano If an object is not a piano then it is not an accordion Therefore if an object is a musical instrument then it is an accordion	Valid	1.675	2.00
If a person is a girl then they are not a son If a person is not a son then they are a daughter Therefore if a person is a daughter then they are a girl	Invalid	4.325	4.57
If an animal is an amphibian then it croaks If an animal croaks then it is a frog Therefore if an animal is a frog then it is an amphibian	Invalid	3.655	4.38
If an animal is a feline then it purrs If an animal purrs then it is a cat Therefore if an animal is a cat then it is a feline	Invalid	3.855	4.67
If an object is a grape then it is a fruit If an object is a fruit then it is a food Therefore if an object is a food then it is a grape	Invalid	4.185	2.06
If an object is an iPod then it is a music player If an object is a music player then it is electronic Therefore if an object is electronic then it is an iPod	Invalid	4.45	1.38

Syllogism	Validity	Average Premise Believability (/5)	Conclusion Believability (/5)
If a food is candy then it is meat If a food is meat then it is a lollipop Therefore if a food is a lollipop then it is candy	Invalid	1.305	4.31
If an animal swims then it is not a fish If an animal is not a fish then it is a dolphin Therefore if an animal is a dolphin then it swims	Invalid	1.585	4.51
If a food is a cookie then it is a vegetable If a food is a vegetable then it is a bean Therefore if a food is a bean then it is a cookie	Invalid	1.295	1.38
If a plant is a weed then it is an oak tree If a plant is an oak tree then it is a flower Therefore if a plant is a flower then it is a weed	Invalid	1.64	1.78
If a person is not a hockey player then he is rich If a person is not rich then he is an athlete Therefore if a person is an athlete then he is a hockey player	Invalid	1.685	1.29

Appendix B: Seven-point FOR Judgment

At the time I responded I felt:

Guessing			Fairly Certain			Certain
1	2	3	4	5	6	7

Appendix C: Six-point FOR Judgment

At the time I responded I felt:

Guessing					Certain
1	2	3	4	5	6