UNIVERSITY OF CALIFORNIA

Santa Barbara

To Detect or Not To Detect:

Dual Process Models and Cognitive Failures

A dissertation submitted in partial satisfaction of the

requirements for the degree of Doctor of Philosophy

in Psychological and Brain Sciences

by

Alexander Benson Swan

Committee in charge:

Professor Russell Revlin, Chair

Professor Mary Hegarty

Professor Richard E. Mayer

Professor Eric R. A. N. Smith

June 2016

The dissertation of Alexander Benson Swan is approved.

_____

Mary Hegarty

_____

Richard E. Mayer

_____

Eric R. A. N. Smith

_____

Russell Revlin, Committee Chair

May 2016

To Detect or Not To Detect:

Dual Process Models and Cognitive Failures

by

Alexander Benson Swan

ACKNOWLEDGEMENTS

There are several people I wish to acknowledge here, without whom the creation of this manuscript and the containment of my sanity would not be possible.

First, this manuscript is dedicated to my wife, Astrid, and my son, Oliver, for their undying support for me in every aspect of my life.

Second, I wish to acknowledge the support and guidance of my doctoral advisor, Dr. Russell Revlin. The research explored and discussed in this manuscript would not have been possible had he not allowed me to forge my own research path. His research insightfulness and pragmatism kept me from traversing too far down the rabbit hole. I am grateful for his patience with my writing and my desire to focus my graduate efforts on teaching. Perhaps I am most grateful that he decided to take one last graduate student, giving me the chance to earn this doctorate.

Third, I must thank the three women who allowed me to actually write this manuscript: my sister, Caitlin Cowley, for jumping at the chance to spend time with her nephew so I could devote full days to writing; my mother-in-law, Libby Seipelt, for the same, especially while Astrid was away on a trip; and last, my mother, Barbie Gonzales, for taking two weeks out of her busy life to come spend time with her grandson so I could write and travel. I also thank her for an entire life of believing in me. There will be very little I will be able to do to repay these efforts so that I could complete this manuscript.

Fourth, I acknowledge the Monday afternoon Cognition Labs seminar group, with whom I have had the pleasure to see every week for the last five years. I especially want to thank Celeste Pilegard and Logan Fiorella for being there right from the beginning—it's been a really fun journey each Monday! The list also includes two of my committee members,

Drs. Mary Hegarty and Richard Mayer. I am grateful to have had the opportunity to chat about my research in a friendly and informal environment.

Fifth, I want to acknowledge the efforts of each of my dissertation committee members, Russell Revlin, Mary Hegarty, Richard Mayer, and Eric Smith, for reading this lengthy and complicated manuscript and offering thoughtful comments and suggestions.

Sixth, I am grateful for all the help from my undergraduate research assistants over the last few years. I would not have been able to collect the data needed for this dissertation or analyze some of the data without these eager folks. It has been an immense joy to mentor undergraduates in research methods, in much the same way I've been mentored in my doctoral training.

Last, I want to thank all the friends and family that were not mentioned above. I don't think my sanity would have remained intact without the fun, the camaraderie, the hugs, the commiserating, and the laughter. It is without a doubt that these things are necessary to write a document such as a dissertation. No matter how small or big your contribution was to this experience, you are all wonderful people to whom I am forever grateful.

VITA OF ALEXANDER BENSON SWAN
May 2016

EDUCATION

Expected June 2016	Psychological and Brain Sciences (Cognition, Perception, and Cognitive Neuroscience) – Ph.D., University of California, Santa Barbara • *Advisor*: Dr. Russell Revlin • *Interdisciplinary Emphasis:* Quantitative Methods in Social Sciences (QMSS)
May 2011	Psychology – M.A. (With Distinction), California State University, Northridge • *Advisor:* Dr. Abraham M. Rutchick • *Emphasis:* General Experimental Psychology • *Thesis Title:* Therefore, Socrates is a talking head: Belief bias and syllogistic reasoning errors in cable news
May 2008	Psychology – B.A. (Summa Cum Laude), California State University, Northridge

Certifications

2015	Certificate in College and Undergraduate Teaching (CCUT) University of California, Santa Barbara
2012	Summer Teaching Institute for Associates Certificate (STIA) University of California, Santa Barbara
2010	Human Participants Protection Education for Research Teams National Institutes of Health

TEACHING EXPERIENCE

Instructor of Record

Teaching Associate
Department of Psychological & Brain Sciences, UC Santa Barbara

Winter 2016	Social Cognition
Summer 2015	Lab in Advanced Research Methods
Summer 2012, 2013, 2014, 2015	Introduction to Experimental Psychology
Summer 2013, 2014; Winter 2014	Health Psychology

Teaching Assistantships

Graduate Teaching Assistant
Department of Psychological & Brain Sciences, UC Santa Barbara

Spring 2016	Motivation
Fall 2015	Introduction to Experimental Psychology
Fall 2014	Introduction to Statistics
Spring 2014	Psychology of the Self

Fall 2013, Winter 2015, Spring 2016	Human Memory
Winter-Spring 2012, Spring 2013, 2015	Laboratory in Human Memory & Cognition
Winter 2013	Cognitive Neuroscience
Fall 2012	Advanced Research Methods Laboratory
Summer-Fall 2012	Health Psychology
Fall 2011	Introduction to Psychology

Graduate Teaching Assistant
Psychology Department, California State University, Northridge

Fall 2010	Introduction to Social Psychology
Fall 2009	Statistical Methods in Psychology

Undergraduate Teaching Assistant
Psychology Department, California State University, Northridge

Spring 2008	Statistical Methods in Psychology

HONORS AND AWARDS

May 2016	*Contribution to Excellence in Teaching Award (CETA)*, Department of Psychological & Brain Sciences, University of California, Santa Barbara
2013-2015	*AAAS/Science Program for Excellence in Science*
September 2011	*Psychology Graduate Research Fellowship*, University of California, Santa Barbara
April 2011	*Honorable Mention, William Wilsoncroft Graduate Award*, Department of Psychology, California State University, Northridge
December 2010	*1st Place, 20th Annual Psi Chi Research Competition,* Graduate Division, California State University, Northridge
November 2010	*Master's Thesis Funding Award*, Department of Psychology, California State University, Northridge
Fall 2004 – Spring 2008	*Dean's List*, California State University, Northridge
2005	*National Dean's List*
Spring 2005 – Present	*Phi Theta Kappa Member*

RESEARCH & SCHOLARSHIP

Peer-Reviewed Publications and Conference Proceedings

Swan, A. B., & Revlin, R. (2015). Inhibition failure is mediated by a disposition toward flexible thinking. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2314-2319). Austin, TX: Cognitive Science Society.

Barrett, M. E., **Swan, A. B.**, Mamikonian, A., Ghajoyan, I., Kramarova, O., & Youmans, R. J. (2014). Technology in note taking and assessment: The effects of congruence on student performance. *International Journal of Instruction*, *7*(1), 51-60.

Swan, A. B., Cohen, A., Evans, S. R., & Drescher, B. A. (2013). Influence of taste quality on affective state. *Psi Chi Journal of Psychological Research, 18*(2), 61-66.

Swan, A. B., Chambers, A. Y., & Revlin, R. (2013). Scope of real beliefs in belief revision. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 1414-1419). Austin, TX: Cognitive Science Society.

Invited Talks & Presentations

Swan, A. B. (2016, January). 'How' and 'why' beliefs affect the reasoning process. Paper presented at the Brown Bag Series, Department of Psychology and Sociology, College of St. Scholastica, Duluth, MN.

Swan, A. B. (2015, December). Conflict detection and resolution in dual process thinking: A case for errors as cognitive failures. Paper presented at the Navy Center for Applied Research in Artificial Intelligence, US Naval Research Laboratory, Washington, DC.

Swan, A. B., Chambers, A. Y., & Revlin, R. (2013, August). Scope of real beliefs in belief revision. Paper presented at the 35th Annual Conference of the Cognitive Science Society. Berlin, Germany. [also under Conference Presentations]

Swan, A. B. (2013, April). Revising your beliefs: How the scope of the rule affects your decision-making. Brown Bag Talk Series, Department of Psychology, California State University, Northridge, Spring 2013.

Swan, A. B. (2011). God and moral foundations: Can liberals turn right? Brown Bag Talk Series, Department of Psychology, California State University, Northridge, Spring 2011.

Conference Presentations (*undergraduate author)

Swan, A. B., *Hill, A., & Revlin, R. (2015, November). Is a picture worth a thousand numbers? The effect of base-rate images in a base-rate neglect task. Poster presented at the 56th annual meeting of the Psychonomic Society, Chicago, IL.

Swan, A. B., & Revlin, R. (2015, July). Inhibition failure is mediated by a disposition toward flexible thinking. Poster presented at the 37th Annual Conference of the Cognitive Science Society, Pasadena, CA.

Swan, A. B., & Revlin, R. (2014, November). Knowledge updating and the visual impedance effect. Poster presented at the 55th annual meeting of the Psychonomic Society, Long Beach, CA.

Swan, A. B., Chambers, A. Y., & Revlin, R. (2013, August). Scope of real beliefs in belief revision. Paper presented at the 35th Annual Conference of the Cognitive Science Society. Berlin, Germany.

Swan, A. B. (2013, May). Formal reasoning by liberals and conservatives. Paper presented at the annual Psychological and Brain Sciences Department Mini Convention, Santa Barbara, CA.

Swan, A. B., Revlin, R., & Rutchick, A. M. (2012, May). Working memory and the belief bias effect

in political arguments. Poster presented at the annual meeting of the Association for Psychological Science, Chicago, IL.

Gold, J. M., **Swan, A. B.**, *Fram, S., *Maru, S. Y., & Rutchick, A. M. (2012, January). Moral decision-making and religious priming: Implications for moral foundations theory. Poster presented at the Judgment and Decision-Making Preconference at the annual meeting of the Society of Personality and Social Psychology, San Diego, CA.

Swan, A. B. & Rutchick, A. M. (2011, April). God and moral foundations: Can liberals turn right? Poster presented at the annual meeting of the Western Psychological Association, Los Angeles, CA.

Swan, A. B., Barrett, M. E., Rutchick, A. M., & Slepian, M. L. (2011, April). You are what you drink: Object priming and persistence. Paper presented at the annual meeting of the Western Psychological Association, Los Angeles, CA.

Ghajoyan, I., Mamikonian, A., Barrett, M. E., Kramarova, O., **Swan, A. B.**, & Youmans, R. J. (2011, April). An examination of encoding specificity in a classroom context. Poster presented at the annual meeting of the Western Psychological Association, Los Angeles, CA.

*Jacobson, A., *Babush, M., Barrett, M. E., **Swan, A. B.**, & Rutchick, A. M. (2011, April). Object priming: The effect of ibuprofen on the subjective experience of pain. Paper presented at the annual meeting of the Western Psychological Association, Los Angeles, CA.

*Babush, M., *Jacobson, A., **Swan, A. B**., & Rutchick, A. M. (2011, April). Increasing pain tolerance via subliminal priming. Poster presented at the annual meeting of the Western Psychological Association, Los Angeles, CA.

Swan, A.B. & Rutchick, A. M. (2011, February). Therefore, Socrates is a talking head: Syllogistic reasoning in cable news. Poster presented at the 15th Annual CSUN Student Research and Creative Works Symposium, California State University, Northridge.

Barrett, M. E., **Swan, A. B.**, Rutchick, A. M., & Slepian, M. L. (2011, January). You are what you drink: Object priming, motivation, and self-rated personality. Poster presented at the annual meeting of the Society for Personality and Social Psychology, San Antonio, TX.

Mamikonian, A., **Swan, A. B.**, Kramarova, O., Ghajoyan, I., & Barrett, M. E. (2010, December). In-class technology usage and its effects on scholastic performance. Poster presented at the semi-annual CSUN Psi Chi Research Competition, Northridge, CA.

Swan, A. B. & Rutchick, A. M. (2010, April). Play it backward: Hidden messages in reversed audio. Poster presented at the annual meeting of the Western Psychological Association, Cancun, Mexico.

Rutchick, A. M., *Coleman, S. R., Corral, D., *Ferber, S., & **Swan, A. B.** (2010, April). The pen is mightier than the word: Priming evaluative harshness. Poster presented at the annual meeting of the Western Psychological Association, Cancun, Mexico.

*Cohen, A., ***Swan, A. B.**, *Evans, S. R., *Tinajero-Guerrero, N., & Drescher, B. A. (2008, August).

Influence of taste quality on affective state. Poster presented at the annual meeting of the American Psychological Association, Boston, MA.

Research Positions

September 2011 – Present	Doctoral Student Researcher/Doctoral Candidate Reasoning Lab, Department of Psychological & Brain Sciences, UC Santa Barbara *Advisor*: Russell Revlin, Ph.D.
August 2010 – May 2011	Lab Coordinator Applied Social Psychology Lab, Psychology Department, California State University, Northridge *Advisor*: Abraham M. Rutchick, Ph.D.
September 2009 – May 2011	Graduate Research Assistant Applied Social Psychology Lab, Psychology Department, California State University, Northridge *Advisor*: Abraham M. Rutchick, Ph.D.
June 2010 – May 2011	Graduate Researcher Psi Chi Graduate Research Team, California State University, Northridge *Advisor*: Robert J. Youmans, Ph.D.

SERVICE

University Service

April – September 2014	Social Sciences Peer Facilitator Summer Teaching Institute for Associates (STIA), Office of Instructional Development University of California, Santa Barbara
October 2013 – June 2014	Graduate Student Representative Graduate Affairs Committee, Department of Psychological & Brain Sciences University of California, Santa Barbara
October 2012 – June 2014	Graduate Student Representative Academic Program Review Panel University of California, Santa Barbara
June – August 2012	Research Mentor Summer Research Mentorship Program University of California, Santa Barbara
August 2010 – May 2011	Peer Mentor Psychology Department Graduate Peer Mentor Program

California State University, Northridge

Professional Service

2013 — Ad-hoc Reviewer, Association for Psychological Science Student Caucus (APSSC)

- 2013 RISE Research Award
- 2013 Student Research Award (SRA)

2013 — Ad-hoc Reviewer, *International Journal of Instruction*

Professional Memberships

Society for Teaching of Psychology (APA Division 2)
Psychonomic Society
Cognitive Science Society (CSS)
American Psychological Association (APA); Graduate Students (APAGS)
Psi Chi, National Honor Society of Psychology, California State University, Northridge Chapter

ABSTRACT

To Detect or Not To Detect:

Dual Process Models and Cognitive Failures

by

Alexander Benson Swan

When faced with a decision regarding probability or heuristics, people generally show their bias toward a heuristic, even if it might be the wrong decision, such as on the classic base-rate neglect task (Kahneman & Tversky, 1973). The crucial question is whether people know that they are focused on this bias. Recent dual process theories (DPTs) have incorporated the crucial role of conflict detection and resolution to better explain why people are biased on classic reasoning and judgment tasks. Two recent models, the Logical Intuition Model (De Neys, 2012) and the Three-stages Model (Pennycook, Fugelsang, & Koehler, 2015) suggest that the source of errors and bias can be explained in two distinct ways. The Logical Intuition Model postulates that people are generally efficient and routine conflict detectors and that errors are due to a failure to inhibit an initial, intuitive response. The Three-stages Model claims that detection is imperfect and that the detection mechanism is the main source of errors because people do not recognize that they are making biased decisions. These claims were investigated in a series of three experiments. In Experiment 1, participants completed a modified base-rate neglect task. In Experiment 2, a conditional

reasoning task was added to test whether the claims of the two base-rate neglect models would transfer to a qualitatively different task. In Experiment 3, participants were placed into four different groups where they were given one of two forms of false feedback, true feedback, or no feedback (control) in order to test whether feedback would interact with answer confidence and resultant intuitive or analytic processing correlates (such as response time). Across all three experiments, the Three-stages Model's claim that monitoring/detection failures is the main source of bias on the base-rate neglect task was supported over the Logical Intuition Model's claim of inhibition failures as the major source of bias. Experiments 2 and 3 support the explanation that these two models are task-specific to base-rate neglect, as conditional reasoning behavioral patterns did not support either model fully. Feedback did not have the predicted effects on accuracy, response times, or confidence. Implications of these findings regarding general dual process theory, including the impact of methodological limitations, are discussed. Small modifications to the Three-stages Model are offered to reflect the data presented here.

TABLE OF CONTENTS

Chapter I.

Introduction

Humans are generally poor reasoners, especially when reasoning is based in logic or probability, and many are susceptible to beliefs and biases when making complex, or even simple, decisions. For example, in a recent episode of *Last Week Tonight with John Oliver*, Mr. Oliver described in excruciating detail the central, biased problem with science communication and consumption in our society: too much confirmation bias. Perhaps the summative point came from a clip of TV weatherman Al Roker, who suggested that a person should merely find a scientific study that confirms their beliefs and use it as evidence to justify their behaviors (Oliver & Pennolino, 2016). Of course, this is usually a bad idea, and clearly shows biased judgment and reasoning.

So why are we so poor at reasoning? Perhaps the answer lies in the idea that there is a dual processing of information, and much of the susceptibility toward errors is due to the issues associated with one or both of those processing routes. Dual process theories (DPTs; see Table 1 for a list of acronyms used in this manuscript) have been investigated in cognitive psychology for over 50 years, but the issue extends much further into the history of psychology (Frankish & Evans, 2009). Though they are varied in their application and description, the main distinction in DPTs is that there is fast, heuristic, and automatic processing (Type 1, or T1) and slow, deliberate, and analytic processing (Type 2, or T2) (e.g., De Neys, 2012; Evans & Stanovich, 2013; Evans, 2003, 2007; Frankish & Evans, 2009; Stanovich, 2009). Depending on the scientific investigation, there are essential factors to consider, such as the defining feature that separates the two types of processing, the related processing associated with each type for a given cognitive task, and the progression of

cognitive resources allocated to each type of processing over time (Evans & Stanovich, 2013).

The goal of this dissertation is not to give an answer to each of these overarching issues in DPTs, but to focus on an important few. Recently, two theoretical models have been introduced that attempt to explain perhaps the most important cognitive aspect of DPTs: conflict detection and resolution (De Neys, 2012; Pennycook, Fugelsang, & Koehler, 2015). Thus, the goals of this present set of studies were to (1) investigate these two models with their relevant behavioral predictions regarding conflict detection and resolution, including decision time, accuracy, and *when* and *why* the two routes defined above might be activated or ignored, (2) broaden the scope of these recent conflict detection models and their behavioral predictions from one major judgment/reasoning task to another task that is qualitatively distinct, and (3) to determine if a performance feedback intervention can mitigate the source of this bias and errors on these two tasks.

Historical Background

Although DPTs has been introduced and studied in various subfields in psychology, such as social psychology (Chen & Chaiken, 1999; Petty & Wegener, 1999) and educational psychology (Epstein, 1994), this dissertation focused on reasoning, judgment, and decision-making (JDM; Tversky & Kahneman, 1974). Many of the current theories in DPTs are from a few authors in the reasoning field (e.g., De Neys, 2012; Evans & Stanovich, 2013; Evans, 2003, 2007; Frankish & Evans, 2009; Pennycook et al., 2015; Stanovich, 2009). Some of the early researchers have suggested that reasoning ability is based on the differentiation between implicit processes vs. explicit processes (Evans & Over, 1996), associative processing vs. rule-based processing (Sloman, 1996), or an automatic set of systems vs. an analytic system

(Stanovich, 2011).

More specifically, Evans and Over (1996) argued that a dual process thinking architecture makes sense when viewed through a rationality lens. They claimed that instrumental rationality—achieving one's goals—need not be tied to rules bound within logic or probability, but influenced by a person's belief system. In other words, thinking progresses to achieve goals already contextualized by beliefs and prior knowledge (Frankish & Evans, 2009).

In parallel with Evans and Over (1996), Sloman (1996) was proposing his dual process theory. His paper was influential in instigating further investigations into dual processing, though he focused on the effects of reasoning (Frankish & Evans, 2009), counter to the approach of Evans and Over. The major contribution of Sloman's paper was his argument that there was an associative system that was based on similarity (prior knowledge) and a rule-based system that processed the symbols of logic. Moreover, these two systems operated simultaneously, whereby the rule-based system could inhibit the associative system if a conflict of response outputs was found.

Last, Stanovich (2009, 2011) added major nomenclature to the dual process debate (i.e., "System 1" and "System 2"), but more importantly, added individual differences hypotheses to the exploration. His idea was that T2 was linked to differences in general intelligence, whereby students with higher cognitive ability tend to do better on reasoning and judgment tasks (Frankish & Evans, 2009). Though the above researchers approached reasoning from different angles, it appears that there are similarities how they describe T1 and T2 processing. It is clear there is strong overlap in the nomenclature used, as well as the synonymic nature of the theories.

Table 2 shows the related processing correlates for T1 and T2 that have been found in associated DPTs research (table adapted from Evans, 2008). T1 processing shows a relationship of fast and automatic thinking. Heuristic thinking, defined as an associative solution based on previous experience (e.g., a rule-of-thumb or an educated guess) when an exhaustive search is impractical or impossible, is also commonly used to identify T1. However, caution must be used here. In the realm of normative logical reasoning, a heuristic tends to be a naughty word, usually indicating that a person has gotten the reasoning problem wrong. This is not always the case; a reasoner can also arrive at a wrong decision using T2 processing (Evans, 2012; Stanovich, 2011; Thompson, Prowse Turner, & Pennycook, 2011). As such, *heuristic* in the DPTs sense merely reflects a faster route of processing and *shortcuts* based heavily on probability and expected outcomes from experience. Indeed, heuristics are used based on the fact that they return a correct answer more often than chance expectation. Due to the automaticity of T1 processing, it is argued to be the *default* state, and T2 is only engaged when it is needed, acting as a shallow monitor to the former (this is described as *default interventionism* by Evans, 2007). By this logic, it appears that T1 guides the human experience on a daily basis. However, because following heuristics and shortcuts are not always the optimal decision, T2 should excel when the situation demands further reflection.

Table 2 also shows T2 processing correlates. Overall, it describes that T2 thinking requires effort, is deliberate, and is empirically slower in decision-making (or rather, slower decisions are deemed to be caused by the activation of mental effort). Evans (2009) argues that the defining feature in DPTs is the engagement of working memory (WM), which is the act of engaging T2 processing. A simple problem might not necessarily need WM to achieve

a decision, which would suggest T1 thinking is sufficient. However, any instance where WM is loaded, the person is conscious of this load, which suggests that activation of WM components is the active engagement of T2 thinking (De Neys, 2006). Stanovich (2011) argues for a defining feature on a more conceptual level: the ability to cognitively decouple. Cognitive decoupling is an umbrella term associated with higher order thinking, such as counterfactual thinking, hypothetical thinking, and reflective thinking. A person's ability to think about situations that are not presently occurring is the essence of cognitive decoupling and the engagement of T2 thinking. Both of these defining features strongly suggest the separation of T1 and T2 processing (Evans & Stanovich, 2013).

Specifically, within the general cognitive function of reasoning, DPTs have included efforts to fine-tune T1 and T2 processing correlates. Initially, the distinction between two processing routes was investigated on the Wason selection task (Evans, 2008). On this task, participants are given a conditional statement and four corresponding cards. Two of the cards present the antecedent of the conditional and the absence of the antecedent, while the other two cards represent the presence of the consequent the absence of the consequent. Responses on this task were attributed to either a basic matching bias (matching the words in the conditional statement to the cards presented) or rationalization (a think-aloud protocol was utilized here to uncover the reasons why choices were made). This was further elaborated into a heuristic-analytic theory (Evans, 2008), where a heuristic answer is generally a preconscious judgment, whereas an analytic judgment involves working memory (WM), where previous knowledge is retrieved from long-term memory and examined more closely within the context of the reasoning problem or situation. In this theory, Evans draws a distinction between incorrect answers based on heuristics (such a matching bias) and the

utilization of WM processing to recognize that the correct answer on the task involves confirmation and falsification of a given rule. He argues that the use of WM is the instantiation of T2 processing, and therefore an analytic process.

The majority of the historical investigations of dual processing have been focused on the belief bias effect. It was first proposed as a dual process effect by Evans and colleagues (Evans, Barston, & Pollard, 1983). Briefly, the belief bias is the result of errors attributed to the mismatch in people's prior beliefs and the normative deductive logicality of a syllogism. Essentially, people tend to accept conclusions that are believable and reject conclusions that are unbelievable. In a conventional belief bias paradigm, participants are asked to evaluate the conclusion of a syllogism (three statements, where the first two are premises to the argument and the third is the conclusion) on whether it is logically valid or invalid (usually these are deductive arguments). The researcher manipulates these problems to be valid or invalid as well as whether the conclusion is believable or unbelievable. The mismatch, or conflict, occurs when a believable argument is actually invalid or an unbelievable argument is actually valid.

The robust findings show that the errors are a result of a strong focus on belief in the content of the conclusion, or what would be outputted by the heuristic route (T1). A person believes a conclusion to be true, even if it might be an invalid conclusion, and determines that the argument is valid. Upon further reflection, however, this person would incorporate the rules of logic into decision-making (activation of more analytic thinking, T2) and choose that the conclusion is invalid regardless of what his or her previous beliefs suggest. These data suggest that deductive reasoning is not black-and-white thinking, but that there are situations where heuristics (beliefs or other previous experiences) play a role in the decision-

making process.

This departure from reasoning as an all-or-nothing cognitive function is seated within a related debate about human rationality and dual processing. Reasoning paradigms generally have a *right* answer (so-called *normative* answers; Evans & Over, 1996; Evans, 2014; Frankish & Evans, 2009). For example, in deductive syllogistic reasoning, there are certain figures and moods that yield valid conclusions and there are other figures and moods that yield invalid conclusions (Evans et al., 1983). Due to this construction, and invention by humans, there is a normative answer that would be deemed correct. Based on performance on a deductive syllogistic reasoning task, a researcher can make a judgment regarding the reasoning ability of a group of participants. The researcher would invariably suggest that low performing people have poor reasoning skills, and possibly suggest that these people are irrational in some way. This may not be the case. Research has shown that, while people do make reasoning errors using T1 thinking, errors can be made when T2 thinking is firmly engaged (e.g., Thompson et al., 2011). Thompson and colleagues (Thompson et al., 2011) found that if you ask a person to make a quick decision, but allow them to rethink and reflect on their answer choice, this could lead to errors. If normative reasoning is to be taken as the golden rule, then it seems most (if not all) humans are irrational. Of course, Cohen (1981) has argued that this cannot be the case, no matter what experimental evidence in reasoning shows (see also Oaksford & Chater, 2001). While the current dissertation utilized the methodology of a normative reasoning task, it is not its intent to prescribe the idea that performance on the task has anything to say about the broader issue of human rationality, but rather speak to a lesser evil: bias.

Last, DPTs is not a single, unified theory of thinking or reasoning. Rather, it is a class

of theories, each with a unique contribution to the discussion (Frankish & Evans, 2009), or could even be regarded as a metatheory (Pennycook et al., 2015). Due to this piecemeal approach, DPTs are not without detractors (Keren, 2013; Kruglanski & Gigerenzer, 2011; Kruglanski, 2013; Osman, 2004). The overarching criticism is that DPTs are not parsimonious when compared with unimodal processing models where thinking and processing happen on a continuum (e.g., Osman, 2004). While many of the criticisms exist on a semantics/terminology level (Evans & Stanovich, 2013), there is one complaint by DPTs critics that is particularly compelling for the future of the class of theories. They argue that there is no straightforward model that accounts for when a heuristic answer conflicts with that of an analytical answer on a simple reasoning problem, or why a person would need a system that regulates these conflicts (e.g., Kruglanski & Gigerenzer, 2011). In response to this, researchers have recently modified their existing models and theories to incorporate the mechanism of *conflict detection and resolution* (Evans, 2007). *Conflict detection* can be broadly defined as the ability of the reasoner to recognize that a problem or context cues multiple, conflicting response outputs; *conflict resolution* can be broadly defined as the response output/behavior/judgment/solution (note that this resolution says nothing about the optimal or normative solution to a given problem; De Neys & Glumicic, 2008; Evans, 2007; Evans & Frankish, 2009). The present dissertation attempts to extend the discussion of conflict detection/resolution in DPTs and specifically test hypotheses and prediction of two recent models described below.

Base-rate Neglect and Dual Process

Though DPTs in reasoning have been primarily investigated with the Wason selection task and the belief bias effect (Evans, 2008; Evans & Stanovich, 2013), another research

avenue rich with heuristics has been explored. The creation of conflict detection and resolution models has utilized the *base-rate neglect* task for dual process exploration. Sometimes referred to as the "lawyer-engineer" problem (Tversky & Kahneman, 1974), this task uses the tendency of people to ignore numerical information (base-rates) when making judgments about group membership. It is an effect of the representativeness heuristic (Tversky & Kahneman, 1974). For example, imagine I sampled 100 people about their occupation and collected personality and trait descriptions of each person. Then imagine I tallied the occupations and determined that 70 were lawyers and 30 were engineers. I randomly select one individual from the sample and recite a description that sounds like a (stereo)typical engineer. I then ask a willing participant to determine to which group this person belongs (i.e., what group membership is *most likely*?). The neglect arises if the participant decides to ignore that the lawyer group had larger base-rate in my sample. This is due to incongruity of the base-rates and the description of the person (incongruent or conflict problems). If the breakdown of the base-rates were switched, where I had 70 engineers and 30 lawyers, but the description of the randomly selected individual remained the same, participants tend to choose the larger group category quickly and with very little error. This is due to the congruity between the probability that 70 engineers having the larger chance of selection and description of my (stereo)typical engineer (congruent or nonconflict problems).

The base-rate neglect task was first applied to DPTs by De Neys and Glumicic (2008). They decided it was a good task to get a high contrast in errors, considering that most people perform poorly when the base-rates and description are incongruous (Tversky & Kahneman, 1974). Following the base-rate judgment task, participants were asked to recall the base-rate information from a subset of the problems they had just seen. Amount correct

(accuracy) and response time (RT) were also assessed as dependent variables.

A number of interesting predictions and findings came out of this study, since DPTs had not yet been applied to many of the heuristic literature. First, De Neys and Glumicic (2008) predicted that the answer cued by a stereotypical description would be chosen more often on problems where the base-rates and descriptions are incongruent than when they are congruent. This is perhaps the most intuitive prediction, and it represents the tendency of the participants to rely upon T1 processing and not entertain the base-rates or what those numbers mean probabilistically (a T2 executable). Second, they argued that a slower response time would be indicative of slower processing times, or the activation of analytic thinking and T2 deliberation. Moreover, the longest response time would reflect the idea that the incongruence between the correct answer and the more salient answer had been detected and the proper inhibition of that response was necessary to arrive at the correct answer. Third, proponents of DPTs had up until this point argued that recognizing the incongruence would be explicitly recalled and would be apparent from a think-aloud protocol. In other words, while solving these base-rate problems, participants would note that the base-rates and stereotype description did not match and make the corresponding adjustment to their judgments while saying that was the reason for the adjustment. Last, the recall task would add additional evidence that the incongruence would be recognized and that base-rates would be processed at a deeper level, benefiting recall.

De Neys and Glumicic (2008) found support for all of these predictions, excluding the explicit recall, contrasting with the viewpoints of Tversky and Kahneman (1974). First, De Neys and Glumicic concluded that a conflict detection system operates constantly, able to activate when needed, and this operation is efficient (fast processing) and routine (normal).

Second, while this conflict detection system works well, resolving the conflict (inhibiting the compelling T1 response) separates individual into bias-susceptible or bias-resistant groupings. De Neys ultimately combined these observations into a single predictive model (De Neys, 2012).

These conclusions have implications about the processing of the information, when that information is processed, and how it connects the two largest pieces of DPTs: T1 and T2 processing. This has increased the conflict detection/monitoring mechanism's role within recent and current DPTs investigations. Other researchers have more recently investigated base-rate neglect and have supported and advanced the initial De Neys and Glumicic (2008) conclusions (Pennycook, Fugelsang, & Koehler, 2012; Pennycook, Trippas, Handley, & Thompson, 2014; Pennycook & Thompson, 2012). Pennycook et al. (2012) replicated De Neys and Glumicic (2008) with the extreme base-rate values used (e.g., 997 vs. 3), but could not replicate the effects when using the original base-rates (described above, 70 vs. 30). However, they did conclude that conflict detection as a monitoring system does operate at some level within T1, even before T2 is engaged. Pennycook and Thompson (2012) further argued that a person's utilization of base-rates is not within the domain of T2, but under the purview of T1. Their results showed that use of the base-rates within a problem (when participants are asked to respond intuitively) is as effortless as incorporating the stereotypic information (which has always been assumed to be a T1 process; see also Pennycook et al., 2014). In the next section, I describe the framework under which the conflict detection mechanism operates, i.e., how the mechanism is proposed to work.

Conflict Detection and Resolution Mechanism in Dual Process

An open investigation in the DPTs literature is the operation of the mechanism of

conflict detection and subsequent resolution (correct or incorrect). Broadly, conflict arises when a solution to a given problem can be answered by an intuition (e.g., a stereotype) or an alternative, possibly logic-based response (e.g., probability) and they are different (De Neys, 2012, 2014a). A person faced with this sort of judgment has an implicit choice to make with respect to a decision. The intuitive response could be chosen, reflecting a desire to stick with the T1 output. Alternatively, the normative response could be chosen, reflecting an initial, salient (correct) response generated by T1 or a T2 output, arrived at by deliberation of the conflict. Represented this way, it appears that T2 only enters the equation when T1 has made an error. The reason for the error could be due to numerous apparent shortcomings of T1 processing, such as errant heuristics that do not apply to a given situation, processing that was too fast or incomplete to fully address the situation, overconfidence, or that an intuitive solution is patently false (De Neys & Glumicic, 2008; Evans & Curtis-Holmes, 2005; Thompson et al., 2011). Of course, why wouldn't a person want to recruit consciousness and WM into a situation in order to arrive at a correct or, at the very least, thought-out solution? This situation is where a well-defined conflict detection system is necessary in new and existing theories (Pennycook et al., 2015).

A central question to the conflict monitoring mechanism is how it works. Researchers have postulated various mechanisms that monitor the actions of T1 and T2 thinking and act as a switch (De Neys, 2012; Evans, 2009; Stanovich, 2011). Recently, Evans (2009) has argued for a "Type 3" processing, which is only activated when there is conflict between T1 and T2. He has called the transitioning between the two main processing types *default interventionism* (Evans, 2007). Evans argues that at any given time, a person is usually utilizing automatic processing (T1), and only when there is an instance of conflict, T2 may

"intervene". The watchful sentry or shallow monitor in this model has the sole purpose to facilitate the transition from T1 to T2 in case of conflicting response outputs. Figure 1, Panel A illustrates the transitioning from T1 thinking (intuitive) and T2 thinking (deliberate), conceptualized for these serialized models. Another serial model that is of note is Stanovich's (Stanovich & Toplak, 2012; Stanovich, 2009, 2011) tripartite model. He refers to T1 as reflexive processing and T2 as reflective processing, with the detector of conflict as a separate algorithmic mind, which handles the T2 processes of cognitive decoupling and serial associative processing. However, these serial models have one glaring deficiency: how is conflict detected in the first place? If there is a conflict, but T1 is the only process engaged, how has T2 detected the conflict, let alone even attempted to resolve it? Recent data does not support this model either: If people are spending more time on problems, what is the nature of this processing? It cannot be T1 as a heuristic system, and if it is T2, what brought about its activation? A serial framework cannot answer these questions. Purely serial models suffer from a tautological impasse.

Another early postulated apparatus for dual processing is a parallel structure (e.g., Epstein, 1994; Sloman, 1996). For example, Sloman's (1996) model suggests that associative processing and rule-based processing operate simultaneously (B, Figure 1), but in the 20 years since the introduction of that model, evidence shows that this is not the case (De Neys, 2014a), let alone too cognitively wasteful (since the model argues that rule-based processing is the more demanding, effortful processing). This creates a situation whereby both systems are always activated, as constant cognitive effort. De Neys and Glumicic (2008) eloquently put this inefficiency as a "viola[tion] of the principle of cognitive economy" (p. 1277). Moreover, if T2 is always engaged, there is little advantage to a heuristic system being in

place (De Neys, 2012). However, the most compelling piece of evidence regarding the parallel structure of dual processing is that it would maintain a conflict monitoring system operating at near perfect efficiency because conflict between the two processing routes is almost always detected (De Neys & Glumicic, 2008). Evans (2007) contradicts this viewpoint by arguing that in parallel processing models, data fit as well as a model where conflict is avoided in the first place. Thus, solely parallel processing seems inadequate to explain recent evidence of reasoning task performance.

Neurophysiological evidence for conflict detection. There is a small body of neurophysiological evidence to support the role of a conflict detector in DPTs. Though a small focus in the DPTs literature, the evidence converges to a couple of brain regions. The anterior cingulate cortex (ACC) is the central brain region that is involved in conflicts of information processing (Botvinick, Cohen, & Carter, 2004). More specifically, the literature reflects that processing in the ACC's functioning is a combination of conflict monitoring and the triggering of cognitive control to modify subsequent performance. Tasks that involve overriding cued or compelling responses are familiar in psychological treatments and aim to test the monitoring and control functions of the ACC. These include Stroop tests, flanker tasks, the Simon task, global-local tasks, and go/no-go tasks. These tasks all share one central idea: the participant must override the overwhelming incorrect choice (e.g., the color of the ink of a displayed word, such as the word *red* in brown ink) for the correct choice (e.g., the actual word written, *red*). The evidence suggests that ACC activation in this way is generally associated with response selection; in other words, it is activated after other regions have processed the information and a behavioral action must be made (Botvinick et al., 2004).

Based on this work that shows conflict activates the ACC, De Neys and colleagues

(De Neys, Vartanian, & Goel, 2008) gave participants base-rate neglect problems while in a functional magnetic resonance imaging (fMRI) scanner to determine if the ACC activates when a person reads problems where the base-rates and the stereotype information are incongruent (much like the classic lawyer-engineer problem; Tversky & Kahneman, 1974). They did not observe differential activation in the ACC, but did note that the right lateral prefrontal cortex (RLPFC) activated on conflict problems. The RLPFC is associated with inhibition processing in reasoning (Goel, Buchel, Frith, & Dolan, 2000; Goel & Dolan, 2003). While the ACC is still argued to be involved, it is possible the base-rate neglect paradigm is not sensitive enough for differential activation. Another possibility is that differential activation in the ACC reflects conscious awareness of conflict, such as "oh, that was a no-go trial, I should not have pressed the button." More work is needed in the neural correlates of conflict detection to make any reasonable conclusion about the ACC's involvement.

In support of the RLPFC participating in conflict detection and resolution, Stollstorff, Vartanian, and Goel (2012) found additional evidence for the RLPFC's involvement as an inhibitory activation on a deductive reasoning task. They found consistent activation of the RLPFC under conditions where the conclusion of the syllogistic argument has a belief-logic conflict. In addition, when the level of conflict in the entire argument was high (i.e., the premises conflicted with belief at varying levels of interference), the RLPFC was engaged, and this activation was above and beyond the activation noted for the conclusion only. The authors argue that the role of this region is to inhibit competing responses, and in this particular case, prior beliefs. Behavioral work by De Neys and Franssens (2009) corroborates this inhibition hypothesis through a series of experiments using various reasoning and

cognitive tasks.

Other neurophysiological evidence (De Neys et al., 2010) in the realm of skin conductance responses has shown that when people solve classic syllogisms that have incongruent information (a belief-logic conflict), an increase in skin conductivity is noted, which is argued to be associated with unconscious activation of a conflict detection system. An autonomic response of the sympathetic nervous system was observed when individuals read a simple syllogism, one in which deduction (validity) suggested one answer (the correct one) and prior beliefs suggested the other answer (the wrong one). The autonomic arousal under conflict suggests detection is an unconscious effort, related to other neural mechanisms of conflict detection.

It is unclear from these limited studies that conflict detection and resolution is primarily the role of the ACC or the RLPFC. Studies that show ACC activations are indicative of conflict detection have not been extended to reasoning studies (and dual process investigations) with a comprehensive approach. Additionally, the influence of the RLPFC has only been shown within belief bias studies and has also not been extended in a comprehensive way. If either brain region is to be attributed to DPTs and the conflict monitoring mechanism, more work is needed. These investigations will be left to future work, as these questions are beyond the current scope of this dissertation. However, cognitively, the current work does investigate the need for a greater specification of the mechanism itself and the situations under which it operates, which the next two models attempt to provide.

Conflict Detection and Logical Intuitions

The first model to be described and tested in this dissertation is De Neys' Logical

Intuition Model (LIM; De Neys, 2012, 2014a). The LIM combines the two approaches discussed above, serial and parallel processing, into a hybrid model (C, Figure 1; De Neys, 2012). De Neys has argued the case for logical intuitions, or rather, simple logical processing that is automatic, due to the conflict monitoring system working at an efficient level (De Neys & Glumicic, 2008). For example, most of the research in conditional reasoning and categorical reasoning has shown that reasoners are adept and quick at completing modus ponens inferences (e.g., De Neys, 2006). The data from this study suggest that this inference is actually a logical intuition and it is such a common occurrence that historically it was formalized into a logical rule. According to the model proffered by De Neys, heuristic intuitions and logical intuitions operate in parallel (the latter might also be considered "shallow" analytic processing; De Neys & Glumicic, 2008; Pennycook & Thompson, 2012) within T1 processing, and only if these two streams of intuitive processing are in conflict (producing different response outputs) would T2 processing be engaged. This is less cognitively demanding than a purely parallel system of T1 and T2 (De Neys, 2012). This is perhaps the crucial difference between the LIM and the previous two processing accounts. The time of conflict is clearly defined as the moment when the intuitive streams of T1 are at odds, reflecting two or more responses. Since T2 is regarded as the analytical processing type, it can now analyze the conflict of the two intuitions and proceed from there. However, engagement of T2 after the onset of conflict is optional—a person can choose to rely on the stronger output of T1 (Thompson et al., 2011). The seriality of T1-T2 is only instantiated in times of response conflict; if there is no such conflict, T2 remains inactive. As a result, this does not necessarily mean that a correct answer to a judgment or reasoning task would be achieved, as previously noted.

Conceptually, if there is a distinction of T1 and T2 in the cognitive architecture, then a mechanism for detection and resolution of conflict must be an integral part of the system. Previous models had failed to capture this important piece, which has engendered skepticism (Keren, 2013; Kruglanski & Gigerenzer, 2011; Osman, 2004); the revised LIM described above places this mechanism as the integral part of model.

Conflict Detection and the Three-Stages

Pennycook, Fugelsang, and Koehler (2015) described a Three-stages Model (TSM) of dual process and conflict detection, and this is the most recent model in the conflict detection debate (and the second to be specifically addressed/tested in this dissertation). The authors argue that the LIM merely describes conflict detection, and by extension, successful detection, but fails to describe or delineate differences in T2 processing quality. This latter point is the central theme of the TSM model: T2 processing, as an imperfect analytic processing system, should have measurable differences in quality depending on the judgment or reasoning task.

The three stages of the model are displayed in Figure 2. The stages have familiar undertones to the LIM. The first Stage is purely T1, which generates a response upon an initial reading of problem or situation. However, much like the LIM, a person can generate multiple initial responses, and these initial responses are not restricted by the model (so they could be a heuristic intuition, such as a stereotype, or a logical intuition, such as modus ponens). The time to make a decision in Stage 1 is characterized in milliseconds. Stage 2 is the conflict monitoring stage: if the problem or situation contains a conflict in outputs, then Stage 3 is entered and T2 is engaged. This monitoring process is also characterized in milliseconds. If there is no conflict, then Stage 3 is merely the output of the initial response

that is most salient.

In all cases, Stage 3 is T2 engagement and reflects response output (the decision made on the task). The authors make two qualitative distinctions in this final stage: *rationalization* or *cognitive decoupling*. If the initial salient response (let's say the stereotype response of base-rate neglect problem) is ultimately chosen, then the person is said to have *rationalized*: reasons for justification of this response are made (even after successful conflict detection), even if it is not the normative response (perhaps an effortful, belief-based response; see Handley & Trippas, 2015). Alternatively, successful conflict resolution in this model and in this Stage is a process labeled *cognitive decoupling*. This is thought by some to be the defining feature of T2 (e.g., Evans & Stanovich, 2013); the TSM refines this point by only ascribing the ability to decouple as successful detection *and* resolution. This latter point is tricky, because the broader definition of cognitive decoupling includes the ability to rationalize. Taken together, this stage is reflected by slower responding, with final decisions taking seconds of processing, rather than milliseconds in the first two stages.

Testing the Models

The specific goals for testing these two models are two-fold: (1) directly testing the model predictions that these two models suggest, which has not been previously attempted, and (2) extend the predictions beyond a base-rate neglect paradigm. First, while the TSM was developed as a response to the LIM, it did not directly test the hypotheses associated with the model. This is a major oversight if a distinction between the two models was desired. Second, and more importantly, both models have been developed within the base-rate neglect task. While I also utilized this task for my testing, it is essential that these two models be subjected to testing within different judgment or reasoning tasks (as outlined by the boundary

conditions in De Neys, 2014a, and a position abdicated by Pennycook et al., 2015). Therefore, a major part of this dissertation focused on the outcomes of not only the base-rate neglect task, but also a conditional reasoning task (Thompson, 1994). These two tasks are qualitatively different and are adequate candidates for contrasts for these two models, as well as DPTs in general, since the demarcation between probability and representativeness cannot be their primary focus. Manipulating logical structure (validity) with believability is an essential empirical exercise in DPTs as was described above (e.g., De Neys & Franssens, 2009; Evans et al., 1983).

While the authors of the LIM and the TSM might argue that their models are distinct, there are more similarities than differences. However, these points can be tested empirically. First, the LIM states that people are generally efficient and routine conflict detectors (De Neys, 2014a); that is, in instances of conflict, people will at least implicitly recognize that the problem's information cues to multiple initial answers. This can be easily shown by utilizing nonconflict problems as a baseline for decision time and then subtracting it from decision times of conflict problem errors (Pennycook et al., 2015). This is because nonconflict problems have cues that point to the same decision; for example, in a base-rate neglect problem, the larger group base-rate and the stereotype description both cue the same answer, requiring minimal processing regardless of strategy and essentially no reason for T2 engagement. Thus, if people are good conflict detectors, then they should have reliable positive time differences between nonconflict problems and incorrect conflict problem responses, regardless of how often a person chooses the stereotype answer or the base-rate answer on conflict problems. The LIM would predict this outcome. In contrast, the TSM would predict that poor performers do not have reliable conflict detection, and therefore a

positive relationship between performance and response time differences would be observed. In either case, both the LIM and TSM suggest that conflict detection is the integral part of the model, so these predictions are crucial to describing the cognitive architecture of dual processing.

An additional prediction that tests both models is the effort of T2. In the LIM, T2 engagement is optional after conflict is detected (De Neys, 2012), and the only description of T2 engagement comes in the form of inhibition of the initial intuitive response. Thus, a correct answer on a conflict problem (other than guessing), requires T2 to act as an inhibiting agent, which is measured by longer processing times. In the TSM, this is also the case, but a distinction is made with decoupling and rationalizing when T2 is engaged. This is because a person could spend more decision time on a conflict problem but still get it wrong (Pennycook et al., 2015). However, there would be a time difference between rationalizing and decoupling, whereby decoupling is ultimately the longest decisions times (and necessarily correct answers). In both of these cases, accuracy and response times can test these two pieces.

The broad predictions for these additional T2 postulates are the following: (1) Inhibition effects can be described as the response time difference between correct and incorrect responses on conflict problems. Inhibition is a function of T2 processing, and so this would increase overall decision time between these two responses. A positive relationship between accuracy on conflict problems and this response time difference is expected, as more base-rate choices should engender longer greater inhibition response time differences. This inhibition effect is the hallmark of the LIM. (2) The decoupling effect prediction can be described as the response time difference between correct conflict

responses from the nonconflict baseline, which represents the progression of successful conflict detection to successful conflict resolution. The relationship between accuracy on conflict problems and this response time difference should be negative, as participants who choose the base-rate less often must decouple in order to arrive at the base-rate answer. This response time difference should decrease as base-rates choices increase. This decoupling explanation is the hallmark of the TSM.

Conflict Detection and Resolution Errors: Individual Differences

If the LIM/TSM is the cognitive framework of dual processing, what is the behavioral component that is needed to complete the psychological picture? More specifically, what are the individual differences that can be described within and between the LIM and the TSM? Behavioral evidence shows that reasoners are not perfect and errors occur, so what about a person makes these errors occur? De Neys and Bonnefon (2013) offer three accounts for why errors might occur on classic reasoning problems to untangle the occurrence of conflict and resolution in DPTs. These process accounts each have different hypothesized behavioral outcomes and have a relationship with the neural correlates research described above. These accounts broadly incorporate findings from various reasoning tasks, including the findings from base-rate neglect and syllogistic reasoning paradigms. These accounts are described in detail below.

Storage failure. First, a reasoner may commit reasoning errors due to storage failure. Essentially, a reasoner has failed to store or learn the proper formal reasoning knowledge or strategies. In this case, the reasoner would have only the heuristic knowledge of the situation (or a similar situation) to guide them through the reasoning process. Therefore, errors cannot be attributed to the conflict monitoring system, because there is no prior knowledge of the

situation to be in conflict with an intuitive answer (T1 processing). Here the error occurs early in the processing of the situation or reasoning problem and further processing (T2) is unlikely when the decision is made (De Neys & Bonnefon, 2013). Stanovich (2009) argues that a storage failure can describe individual differences data, especially on more difficult reasoning tasks. A storage failure behavioral pattern would likely show a high error rate on conflict problems, coupled with relatively short response times. This latter measure could reflect guessing or the urge to comply with the intuitive answer. The type of failure can be used to show that the differences between a biased and an unbiased reasoner diverge early in the reasoning process.

Monitoring failure. Second, a reasoner might be biased or make errors because there is a monitoring failure. In this account, a person has the necessary formal knowledge, but is unaware that the situation or problem demands this particular knowledge, and so there is a transfer disconnection leading to favoring the heuristic answer. This is a failure of detecting the conflict of a heuristically-cued answer vs. a formally-cued answer. The TSM and LIM diverge on this point: the TSM postulates that conflict detection is an individual difference, and the monitoring failure defined by De Neys and Bonnefon (2013) is synonymous with a failure to detect conflict (Pennycook et al., 2015).

This has also been described as a metacognitive monitoring failure (Thompson, 2009). Thompson and colleagues (Thompson, 2009, 2013; Thompson et al., 2011, 2013; Thompson & Morsanyi, 2012) have described the motivating mechanism of conflict detection and resolution as a *feeling of rightness* (FOR). The FOR describes the process by which an individual remains in T1 thinking, when all indications of a given problem point to engaging T2 thinking. However, a feeling of anything is highly abstract and conceptual.

Thompson and colleagues (Thompson et al., 2011) argue that FOR is maintained in a self-appraisal of *confidence* in a given response. The experience of confidence is an affective one and it is linked to the relative ease by which an answer comes to mind (i.e., *fluency*). Last, T2 engagement is wholly dependent on the strength of the FOR response and fluency of information processing. If FOR is high and an answer comes easily to mind (such as in a base-rate task with a strong stereotype description), then T2 thinking will not likely become engaged because there is little monitoring. A response of this type should be quite quick. If there is any instance of monitoring, this in and of itself should be time consuming, and with decreasing confidence, should increase. Thus, if FOR is low and fluency is equally low, T2 thinking will likely be engaged in order to deliberate on an answer, increasing processing time. To achieve a correct answer with low FOR, even more time is needed than for incorrect answer.

Thompson and colleagues (Thompson, 2009, 2013; Thompson et al., 2011, 2013; Thompson & Morsanyi, 2012) have a number of experimental results to support the FOR account. In one such study (Thompson et al., 2011), participants answered questions during a reasoning task, where each question was followed by a confidence question, which asked the participant to choose whether they were guessing or were certain, on a 1-7 Likert scale. Then they allowed participants to rethink their original answers and possibly make a change to those answers. The researchers found that confidence was negatively associated with the decision not to change answer when given an opportunity (keep the initial answer). In other words, when FOR was low, participants were more likely to change their answers and spend much more time thinking about their answers. Thus, a low FOR signals for T2 engagement and the need for analytic thinking. T2 engagement did not necessarily lead to a correct

answer, however.

Inhibition failure. Third, a reasoner might be biased or make errors because there is an inhibition failure. A person would have the required formal knowledge for the situation, and most of the time, use it. In addition, these people would have the ability to detect and monitor any conflicts between formal answers and heuristic answers. However, the use of formal knowledge and conflict monitoring are theorized to be implicit processes (De Neys & Glumicic, 2008). Essentially, the conflict monitoring processes occur without conscious knowledge (and proceed within milliseconds of processing), but the decision is made on the initial response of T1 (likely the more salient cue; Pennycook, Trippas, et al., 2014). The failure may arise from a number of factors, such as motivation, cognitive resources, failure of deliberation/reflection, or an explicit justification of the why the heuristic solution is correct (rationalization). The behavioral pattern seen by this inhibition failure account would show errors on the conflict problems, at varying levels depending on individual differences, and also shorter response times when compared to correct conflict problems. Correct conflict problems would have the longest response times as it reflects the entire conflict monitoring and resolution process, where conflict is detected and the prepotent (incorrect) response is inhibited (De Neys & Glumicic, 2008).

Recent evidence (De Neys & Franssens, 2009; De Neys & Glumicic, 2008; De Neys, Moyens, & Vansteenwegen, 2010; De Neys, Rossi, & Houdé, 2013; Pennycook & Thompson, 2012; Thompson et al., 2011; Thompson & Morsanyi, 2012) points to the errors of the conflict detection mechanism as a result of a monitoring failure or an inhibition failure.

Incorporation of failures into process models. These failure accounts represent broad descriptions of potential biases. However, they do need to be included within a larger

dual processing framework. Both the LIM and TSM models suggest similar hypotheses for the incorporation of the failures, especially for storage and monitoring failures. There is divergence for inhibition failure between the models.

For storage and monitoring failures, both models agree that conflict would not be detected in people who do poorly across judgment and reasoning tasks (bias-susceptible reasoners). There would be early time divergence between these people and those who do well on judgment and reasoning tasks (bias-resistant reasoners). Bias-susceptibility would produce behavioral responses of low accuracy and fast responses vs. higher accuracy and slower responding marked by bias-resistance. The measurable distinctions within bias-susceptible individuals would be accuracy and confidence on conflict problems: storage failure would reflect a worse performance than monitoring failure, and perhaps less confidence (the trumping measure here, though, is response time, which would be quick).

For inhibition failures, the two models diverge. First, in both cases, late divergence of bias-susceptible reasoners and bias-resistant reasoners occurs: no storage or monitoring failures—bias is simply the result of failing to inhibit. However, the models differ on how this individual difference account and time divergence is described. The LIM claims that inhibition failures are the main reason for errors and the main distinction between sets of people on the base-rate neglect task (De Neys, 2012, 2014a; De Neys & Glumicic, 2008), and recent evidence supports this claim. More importantly, it suggests that conflict detection is usually successful, but in order for a person to be accurate on a given conflict problem, T2 must act as an inhibitor of the salient, compelling (heuristic) initial response. The TSM suggests that successful resolution of a conflict problem is not necessarily inhibition, but T2 engagement and deliberation brought about by cognitive decoupling (Pennycook et al.,

2015). Unsuccessful resolution is the result of T2 engagement, but instead of decoupling, the reasoner rationalizes the initial, compelling response. Clearly, the LIM makes a distinction between T2 engagement: a correct answer on a conflict problem represents T2 (inhibition); an incorrect answer reflects T1 output. The TSM contrasts this view by stating both outputs (correct and incorrect) are the result of T2 and response times should reflect this—one is inhibition (decoupling) and the other not (rationalization).

Are these consequential biases? The final piece for these cognitive failures is if they are consequential to decision-making. Do these failures represent a cognitive framework that is poorly adapted to modern problem-solving (e.g., Nisbett & Wilson, 1977)? If these biases are troublesome, an education focus should be adopted that can indicate when biases can occur on these types of tasks; offering training or feedback might improve performance and aid efforts to reduce biases in general.

In a new line of research, De Neys (2014b) investigated the role of feedback on whether performance on a reasoning task was a monitoring failure or if it is due to an inhibition failure. He compared groups that either got feedback or got no feedback, and found that the typical measures of T2 engagement (accuracy and RT) were modulated by the presence of feedback. Participants benefited by receiving feedback and this was most important on conflict problems. He argued that this was evidence for an inhibition failure account, because participants had already detected the conflict, and the feedback only signaled them to their errors, allowing them to inhibit the prepotent response.

Using a similar feedback manipulation within the tasks the final goal of this dissertation was to mitigate the described cognitive failures. If these failure accounts were based on individual differences, would giving manipulating feedback interact with their

effects? In other words, if a person who tends to make monitoring errors is told that they are doing poorly, regardless of their actual performance, will their confidence decrease and signal Type 2 engagement, effectively extinguishing conflict monitoring errors? If a person is told they're doing well, would the interaction be minimal, indicating errors based on monitoring failures? These are central questions to the roles of confidence, fluency, response time, and failure divergence (early vs. late) and address whether failures such as monitoring can be extinguished by merely reducing participants' confidence. This would ultimately reflect the modal nature (De Neys & Bonnefon, 2013) of the failures, as well as support the process models of LIM/TSM (De Neys, 2012, 2014a; Pennycook et al., 2015).

Thinking dispositions and cognitive ability. A related individual differences investigation in DPTs has centered on thinking dispositions and cognitive ability. Thinking dispositions (or thinking styles) are the motivational component toward utilizing T1 or T2 in domain general or domain specific ways (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014; Pennycook et al., 2015; Svedholm-Häkkinen, 2015; Svedholm & Lindeman, 2013). In other words, how willing is a person to engage in effortful processing in a given context or situation? Research has shown that a person's motivation for complex cognitive activities is associated with the ability to inhibit intuitive responses (Swan & Revlin, 2015) on the Cognitive Reflection Test (CRT; Frederick, 2005); moreover, a person's cognitive flexibility or rigidity (Stanovich & West, 1997) has a pronounced relationship with performance on the base-rate neglect task, among other reasoning tasks and investigations (Pennycook, Cheyne, et al., 2014; Pennycook et al., 2015; Swan & Revlin, 2015; Svedholm-Häkkinen, 2015; Svedholm & Lindeman, 2013). Relatedly, the ability of a person to engage in this type of processing is important (Stanovich & West, 2000). However, the dispositional quality of an

individual trumps their cognitive ability—it's not really a question of *can*, but *want*. Is a person intrinsically motivated to engage T2 on a given reasoning or judgment problem? The crucial part of this question is that it separates the engagement of T2 and a possible normative answer from the initial outputs of T1 (conflict monitoring or initial heuristic/logical intuitions). Utilizing these measures to compliment the investigation of individual differences is fruitful to further identify the biases that occur on various reasoning tasks and how this relates to performance on those tasks.

Present Series of Experiments

The current processing models (LIM and TSM) and failure hypotheses describe certain patterns of behavior, and there is still a debate on the efficiency of conflict detection and the quality of T2 engagement. For example, are errors the result of conflict detection errors (monitoring) or conflict resolution errors (inhibition, decoupling, or rationalization)? The evaluation of the two models will carry through the following set of three related experiments. The first experiment will utilize the base-rate neglect task to test the model predictions of the LIM and the TSM. Participants will be shown 50 base-rate problems, with half of them containing conflicting base-rate and stereotype diagnostic information. The second experiment will expand the investigation from base-rate neglect to conditional reasoning problems (e.g., Thompson, 1994; Thompson et al., 2011), which is a qualitatively different task than base-rate neglect (for a comprehensive review see Thompson, 1994). The addition of conditional reasoning allows for a contrast of logic principles and believability. A comparison of conflict detection/monitoring and metacognitive monitoring will be a central piece of this experiment across both tasks. The third experiment will incorporate a feedback manipulation into the two tasks, designed to either increase reliance on T1 outputs (high

performance feedback) or instigate T2 engagement (low performance feedback). An additional analysis goal of the present studies is the description of group effects, but more importantly, contributing to the individual differences part of this debate.

A number of research questions will be addressed in the present thesis:

Research question 1. Do the data support the LIM framework that states that the conflict detection mechanism is a T1 process or that it is a T2 process, as suggested by the TSM?

Research question 2. Do accuracy, response time, and confidence measures indicate the type of failure with either processing model?

Research question 3. There is converging evidence from two research groups (De Neys, 2012; De Neys & Glumicic, 2008; Pennycook, Cheyne, et al., 2014; Pennycook et al., 2012; Pennycook & Thompson, 2012; Pennycook, Trippas, et al., 2014; Thompson et al., 2011, 2013) regarding base-rate neglect problems and the issues of conflict detection and resolution. However, each of these discussions has viewed the mechanism at the group-level, with minimal discussion of individual differences, and little attention paid to individual-level processing (Pennycook et al., 2015). As such, there is an open question with respect to how often a person experiences conflict detection and modifies their behavior to resolve the conflict (or not). Two related research questions stem from this ID focus:

> **(3a)** Using the failure hypotheses (De Neys & Bonnefon, 2013), can participants be reliably grouped by their performance? In other words, are certain patterns of behaviors indicative of storage failures, monitoring failures, or inhibition failures?
>
> **(3b)** With a large number of trials for a base-rate neglect task or a conditional

reasoning task, performance over time can be assessed. Is it an ongoing mechanism, or after a certain number of experiences with the same type of reasoning problem will a person begin to modify their initial behavior? In other words, is engagement of T2 thinking fleeting for a single problem, or does the conscious activation of T2 thinking engender learning to prevent similar conflicts from T1 thinking? Broadly, are people learning on this task (*the learning hypothesis*)?

Research question 4. What is the nature of FOR in T2 engagement (Thompson et al., 2011)? Findings in support of the FOR argument would indicate that the conflict detection system errors are a result of a monitoring failure.

Research question 5. Is performance on the base-rate neglect task indicative of domain general tendencies that will transfer to a conditional reasoning task?

Research question 6. There is little research on the role of feedback in dual process reasoning. Can providing feedback promote learning on the task and interact with erroneous responding (increase or decrease)?

Additional questions. Additional research questions were investigated after discussion of the impact of the above research questions.

Chapter II.

Experiment 1

Experiment 1 was conducted to test the behavioral predictions inherent in the LIM (De Neys, 2012) and the TSM (Pennycook et al., 2015) dual processing models.

Both models claim that conflict detection is a fast, T1 process, but the LIM postulates that people do it efficiently and routinely, and it is a universal trait; the TSM states that the conflict detection system is error-prone (i.e., prone to monitoring failures). Thompson and others (e.g., Pennycook & Thompson, 2012; Pennycook, Trippas, et al., 2014; Thompson et al., 2011) have found evidence that suggests base-rates are processed within T1 and that conflict is detected within T1, not a stand-alone process between T1 and T2 (as would be suggested by a serial or default interventionist accounts), supporting both models' claims that conflict is borne out of conflicting T1 outputs. RTs coupled with accuracy (i.e., choosing the base-rate answer for either problem type) are crucial to describing the conflict detection and monitoring claims of the two models. It is only the engagement and the quality of T2 that is different between the two models, and RTs in this stage will indicate which model describes the data better.

The LIM claims that T2 engagement inhibits the intuitive response (De Neys, 2012, 2014a, 2014b; De Neys & Glumicic, 2008; Handley & Trippas, 2015) and that the RT difference between correct and incorrect answers on conflict base-rate problems reflects this inhibitory processing. The TSM makes a distinction with correct answers on conflict problems: the longest RTs reflect cognitive decoupling, which has similar qualities to inhibition (Pennycook et al., 2015), but this is reflected by the RT difference between nonconflict responses and correct conflict responses. Cognitive decoupling responses should

take longer than incorrect conflict responses, but only if conflict is detected (i.e., no monitoring failure). According to the TSM, this latter process is described as rationalization. Something about the stereotype is just too compelling to ignore. Thus, the TSM postulates that conflict detection would show a positive relationship between RT and conflict problem accuracy. However, participants with higher accuracy will spend less time *decoupling* than participants with lower accuracy, indicating a negative relationship between processing times in T2 and base-rates responses for more bias-resistant reasoners. This ultimately supports the concept that base-rates are processed in T1 as both models suggest and biased-resistant people can ultimately utilize this information better than bias-susceptible people (De Neys, 2012; De Neys & Glumicic, 2008; Pennycook et al., 2015).

In the case of the conflict monitoring/detection claims of the two models, as well as the cognitive decoupling claim of the TSM, nonconflict RTs are used as a baseline measure to create RT difference scores. The reason RTs on nonconflict problems represent a baseline is because the problems generally have one answer cued by both the base-rates and the stereotypic information in the problem. Historically, accuracy is generally at a ceiling (with some variation if the problem contains less salient or misinterpreted stereotypes; Pennycook et al., 2015). Since a single answer is cued by both sources of information in the problems, then RTs are assumed to be extremely quick beyond an initial reading, acting as starting point for measuring additional processing that is hypothesized to reflect conflict detection and T2 engagement.

These two models identify individual differences regarding the sources of errors on this task: *storage* and *monitoring failures*, which are likely to occur during T1 processing, before the moment of conflict detection (De Neys & Bonnefon, 2013). These sources of

errors would result in poor accuracy on the task, as well as very little RT difference between nonconflict and conflict responses in general. Thus, as they are defined, errors are apparent before any observable T2 engagement. The remaining source of errors, *inhibition failure*, occurs after conflict is detected; any proper resolution would require T2 to act as an inhibitory agent. If the data support storage or monitoring failures, then bias-resistant and bias-susceptible people should diverge rather early in their processing. A person can either detect a conflict or not on a given problem, and additional processing would be required beyond the detection (TSM). In contrast, if the data support inhibition as defined here (LIM), then everyone detects conflicts, and the difference between bias-susceptible and bias-resistant people is the ability to inhibit (giving more processing power to T2. This experiment specifically tests these hypotheses, at both the group level and the individual level by measuring RT differences and comparing them across accuracy on conflict problems.

Furthermore, individual differences can also be described by *thinking dispositions* and *cognitive ability* (Frederick, 2005; Pennycook, Cheyne, et al., 2014; Pennycook, Trippas, et al., 2014; Stanovich & West, 2000; Svedholm-Häkkinen, 2015; Swan & Revlin, 2015). This experiment utilized some of the common indices of these two individual differences measures, including the Cognitive Reflection Test (CRT; Frederick, 2005), the Need for Cognition scale (Cacioppo, Petty, & Kao, 1984), the Actively Open-minded Thinking scale (Stanovich & West, 1997), and SAT scores (Stanovich & West, 2000). Correlations between dispositions toward a desire for engaging in complex cognitive activities (Need for Cognition) and a stronger cognitive flexibility (Actively Open-minded Thinking) are expected to be positive. In accordance with the inhibition failure source of errors, the CRT should be positively correlated with conflict problem accuracy, which suggests that the two

measures both describe a person's ability to inhibit an intuitive, but ultimately incorrect, response. Finally, SAT scores should be positively correlated with conflict problem performance; previous research has shown that SAT scores have been linked to T2 processing and independent of T1 processing (nonconflict problem accuracy; Evans, 2013).

Last, this experiment tested for a learning effect (Mevel et al., 2015). It modified the conventional base-rate neglect task by adding many more problems in order to track progress over time. Base-rate neglect has been studied extensively in the last decade in the conflict detection realm of DPT. However, the methodology rarely has participants complete more than 20 base-rate problems (in total). While there have been extensive-trial studies in base-rate neglect (e.g., De Neys et al., 2008; De Neys et al., 2010; Pennycook et al., 2015), none have discussed participants' progress over time. This approach is novel and may illuminate how reasoners react to multiple problems over time. This methodology additionally tested the viability of introducing feedback in an extensive-trial task.

In the current experiment, the Research Questions 1, 2, and 3 were addressed. The predictions for this experiment were as follows:

Predictions

Behavioral predictions. (1) Performance of participants on conflict problems will be less accurate than performance on nonconflict problems, when setting the "accurate" answer as the base-rate choice on each problem. (2a) *Conflict detection hypothesis*: If there is a conflict detection process that people use, then incorrect responses on conflict problems will be slower than on nonconflict problems, overall, indicating general conflict detection and support for both the LIM and the TSM. (2b) If LIM is supported, there would be a flat relationship between conflict problem accuracy and RT differences (conflict detection index:

incorrect, conflict RT – overall nonconflict RT), suggesting all participants recognize conflict (efficient and routine—ruling out storage and monitoring failures). (2c) If the TSM is supported, this relationship would be positive, suggesting that poor performers do not detect conflict as easily as stronger performers (storage and monitoring failures are viable). (3) *Inhibition failure hypothesis*: correct, conflict problem RTs will be slower than incorrect conflict problem RTs, supporting the LIM; this relationship can also be expressed as a positive relationship between difference RTs for each participant and conflict problem accuracy. (4) *Cognitive decoupling hypothesis*: the RT difference between correct responses on conflict problems and overall nonconflict responses is negatively correlated with conflict problem accuracy, supporting the TSM. This indicates that participants spend more time decoupling on correct answers when those answers are fewer, while participants will spend less time overall generating the base-rate answers when they do it more frequently on the set of conflict problems. Table 3 contains a condensed version of LIM/TSM index predictions (Predictions 2b – 4).

Individual differences predictions. Bias-susceptible vs. bias-resistant individuals will be compared to determine the extent of the above behavioral predictions using analysis methods employed in previous literature (De Neys, 2012; De Neys & Glumicic, 2008; De Neys et al., 2010). Correspondingly, bias-resistant and bias-susceptible individuals can be classified based on behavioral patterns on the task.

Additionally, time series regressions will show the progression of performance across trials. Based on previous work (e.g., De Neys & Glumicic, 2008), the hypothesis is that participants interpret each problem at a time (order is completely randomized) and there will be no appreciable or reliable practice or learning effects.

Method

Participants. Ninety-three psychology undergraduate students (71% female), with a M_{age} = 18.85 years (*SD* = 1.44), participated in this study for partial course credit.

Design and materials. All tasks and measures used in the Experiment and subsequent experiments are detailed in the Appendices.

Base-rate task. The materials used for this study were adapted from De Neys et al., (2008). See Appendix A for the full list. A typical problem appeared like this:

> In a study 1000 people were tested. Among the participants there were 5 men and 995 women. Jo is a randomly chosen participant of this study.
>
> Jo is 23 years old and is finishing a degree in engineering. On Friday nights, Jo likes to go out cruising with friends while listening to loud music and drinking beer.
>
> What is most likely?
> a. Jo is a man
> b. Jo is a woman

This is an example of a conflict problem (base-rate information and diagnostic information are incongruent). The other type of problem that participants answered was nonconflict problems, where the base-rate information and diagnostic information about the randomly selected individual were congruent with one another (swap the numbers for men and women in the above problem). Participants answered 50 total base-rate problems, with 25 conflict problems and 25 nonconflict problems. The neutral condition from De Neys and Glumicic (2008) was not included to maximize the contrast in problems type (neutral problems do not offer stereotypic information). See De Neys and Glumicic for pretesting information regarding strength of stereotype for the problems tested. Stereotypes used varied in content: age, gender, race, job-related groups, and stereotypical human characteristics.

In addition to the congruency of the base-rate information and the diagnostic

information, there were three expressions of extreme base-rates (997 to 3, 996 to 4, and 995 to 5). This was done to vary the presentation of the problems, as well as to force reading of the base-rate information (it could be argued that if this information was the same, it would be ignored).[1] Previous research using this methodology has shown that the extreme base-rates are needed for contrast between the conflict and nonconflict problems (De Neys & Glumicic, 2008; Pennycook et al., 2012) vs. the traditional 70/30 split used in Tversky and Kahneman (1974).

The task began with an overview of the fake survey that participants believed was the basis for the task. For each problem, they were told that a random sample of 1,000 respondents from the survey was selected. Below is the overview:

> In this stage, you will encounter 50 problems regarding a recent comprehensive survey that was conducted in the country. Various pieces of information were gathered from thousands of individuals. Each problem will have a random subset of 1,000 responses.
> For each problem, you will be given a brief description of a randomly selected individual from the subset. **Based on the information provided, it is your task to decide to which group the individual belongs.**

Individual difference measures. In addition to the main base-rate decision-making task, participants completed three individual differences measures in order to get a more complete picture about the decisions that are made on the conflict and nonconflict problems.

Cognitive Reflection Test. The CRT is a behavioral measure that tests a person's ability to inhibit an intuitive response on a word problem (Frederick, 2005). Consider the problem below:

A bat and a ball together cost $1.10. The bat costs $1.00 more than the ball. How

[1] De Neys and Glumicic (2008) performed pretesting on the extreme base-rate values to counter this argument and to vary the numbers in order to draw attention to differences between problems. Post-hoc analysis in their study showed that the small variation did not change performance.

much does the ball cost?

The intuitive answer that is cued in the wording of the problem would be 10 cents, but this answer would be incorrect due to the "more than" phrasing. The correct answer, upon further reflection, would guide the participant to the correct answer of five cents. There are two additional word problems in the Test that have careful phrasing to elicit an intuitive answer that a person would have to inhibit in order to arrive at the correct answer. Using this behavioral method allowed for an objective observation of cognitive reflectivity not achieved by some subjective measures. See Appendix B for the entire problem set.

Thinking disposition questionnaires. Participants were also given two thinking disposition questionnaires. These consisted of 18 items from the Need for Cognition scale (NFC; Cacioppo et al., 1984) and 41 items from the Actively Open-minded Thinking scale (AOT; Stanovich & West, 1997). The NFC asked questions to gauge the propensity of the participant to engage in effortful thinking (T2), such as "I prefer complex to simple problems." Participants rated their agreement with the statements on a five-point Likert scale, where larger values represented a characteristic quality of the individual and smaller values represented an uncharacteristic quality of the individual. The AOT was a composite questionnaire that gauged the cognitive flexibility of an individual. In other words, it measures how willing someone is to engage in effortful processing that has the potential to modify existing beliefs or evaluations. An example of a question from the questionnaire: "Difficulties can usually be overcome by thinking about the problem, rather than through waiting for good fortune." Participants rated their agreement on a six-point Likert scale, where larger values represented stronger agreement with the statement and smaller values represented stronger disagreement with the statement. Each scale had negatively-worded

statements that were then reverse-coded (to prevent response acquiescence). Summing the individual item scores created composite scores for each participant. Greater composite scores reflect a greater propensity to engage in effortful thinking and more flexible thinking. Appendix C includes the set of items for the NFC scale and Appendix D includes the set of items for the AOT scale.

Cognitive ability. In addition to thinking dispositions, which reflect cognitive style (Pennycook, Cheyne, et al., 2014; Pennycook, Trippas, et al., 2014; Stanovich, 2009), a cognitive ability measure used frequently in the cognitive literature (e.g., Stanovich & West, 2000) was gathered from participants, who provided their most recent SAT score. The majority of the undergraduates who participated in the study were from the western United States, so this was the likely standardized test taken prior to coming to college.

Procedure. Participants completed each of the measures at a computer station. Each student was introduced to the study and told that there were four stages to the entire session. Participants first solved the three problems of the CRT. After this, they answered the 50 base-rate problems. The order of these problems was fully randomized. Additionally, the answer was randomized, and it was either presented as the first option or second option (approximately 50% of the problems for each choice and for each problem type). Once the participant finished with those problems, they were given an opportunity to rest. After the short rest period, instructions for the NFC appeared, describing the questionnaire and each of the corresponding scale values. The question order was randomized, and the scale appeared below each question. Upon completion of the NFC, the AOT was presented. Instructions and description of the scale preceded the questions. Question order was randomized and the scale appeared below each question. Finally, participants entered their most recent SAT score (out

of 2400) and some demographic information. Participants were debriefed, thanked, and dismissed.

Results

Main behavioral analyses. Behavioral analyses were conducted to test the four main predictions. The outcome of each prediction occurs in order, with supporting statistical information. In all analyses described below, if a participant did not contribute a value to all cells associated with the specific statistical test, the participant was excluded. Specific exclusions are noted for the corresponding tests.

Prediction (1) stated that nonconflict problems would garner much higher accuracy (defined as choosing the base-rate answer on all problems) than accuracy on conflict problems. This prediction was supported: as shown in Table 4, participants chose the accurate answer on nonconflict problems more often than on the conflict problems, $t(92) = 15.47$, $p < .001$, Cohen's $d = 1.47$. The variance was small in this set of 25 nonconflict problems; three problems could be considered outliers for accuracy values below 2.5 *SD*s.[2] The distribution of problem means for conflict problems was larger, but no problems fell outside of the outlier range. For many of the conflict problems, the stereotype decision was made significantly more often than chance performance (typically low accuracy).

The *conflict detection hypothesis* stated that there would be an increase in RTs for incorrect conflict responses above a baseline nonconflict RT. This simple increase marks the presence of overall conflict detection within the group. Prediction (2a) was supported. As Table 3 reveals, participants took significantly longer to answer conflict problems incorrectly

[2] All analyses were conducted without these 3 problems, but no meaningful changes to accuracy were observed.

than nonconflict problems overall, $t(84) = 5.44$, $p < .001$, $d = .62$.[3] However, this overall difference needs further explanation by incorporating accuracy. Prediction (2b) tests the specific tenet within the LIM that conflict detection is efficient, routine, and ubiquitous.

In contrast, the converse piece of the TSM is also tested, which suggests that participants with numerous stereotype responses on conflict problems are less likely to show conflict detection within their responses (Prediction (2c)). Specifically, RT differences were computed. For each participant, overall average RT on nonconflict problems was subtracted from RT on incorrect conflict responses. This RT difference value is the conflict detection index. Figure 3 illustrates this index along with the other two indices described below. There was a large, positive correlation between conflict performance and the conflict detection index, $r(83) = .70$, $p < .001$, $R^2 = .48$, indicating that as accuracy on the conflict problems increases, so does the difference between incorrect responses on those same conflict problems and nonconflict RT. In other words, people who performed better on the conflict problems had greater RT differences than people who didn't perform well on conflict problems. Moreover, there was a large negative intercept ($b = -2.36$ s, $t(84) = -3.54$, $p < .001$). This outcome reflects support for the TSM (2c) and not the LIM (2b). Some individuals clearly did not express conflict detection within their RTs and it is clearly not routine in all individuals. However, overall, conflict detection does occur, which is at least a corroborating finding for both models.

Prediction (3) tested the *inhibition failure hypothesis*, which is a combination of one of the failure hypotheses and the role of T2 engagement. For the LIM, T2 engagement is

[3] Eight participants were excluded from this analysis for not contributing any incorrect conflict responses.

reflected in longer RTs on correct vs. incorrect responses on conflict problems, which is inferred to represent an inhibition function. For the TSM, inhibition is not necessarily the explanation for correct answers on conflict problems.

Two analyses were conducted to test inhibition failure. First, a paired-samples *t*-test showed that when correct judgments on conflict problems were compared with incorrect judgments, correct judgments had the same overall RT as incorrect judgments, and this difference was not significant ($t(84) = .28$, $p = .78$, one-tailed).[4] From these data, it is difficult to claim support for the LIM and inhibition, as this relationship was not significant, which is contrary to recent support (e.g., De Neys & Glumicic, 2008). This difference could be attributed to the more bias-resistant participants making much quicker base-rate responses as accuracy increased.

Second, an additional index was computed. The inhibition index was computed by subtracting incorrect conflict RTs (stereotype) from correct RTs (base-rate). Positive values on this index indicate longer processing to make the correct decision vs. an incorrect decision, which in turn should have a positive relationship with accuracy. Again, in tandem with the *t*-test above, the LIM prediction was not supported by the data; the inhibition index and conflict problem performance were negatively correlated, $r(83) = -.61$, $p < .001$, $R^2 = .38$. The intercept was also significantly greater than zero ($b = 2.83$ s, $t(84) = 4.55$, $p < .001$), which essentially means that a participant who gets only one conflict problem correct would spend longer than two seconds making that correct response vs. the average time spent processing the incorrect responses. Figure 3 also illustrates the RT difference

[4] The same eight participants as the previous test were excluded from this analysis for not contributing any incorrect conflict responses.

patterns for the inhibition index, with the regression line clearly showing a negative trend.

Finally, the last behavioral prediction of *cognitive decoupling* was tested (4). The cognitive decoupling index was computed by subtracting overall nonconflict RTs from correct conflict RTs. Positive values reflect greater time processing the correct answer on a conflict problem vs. when information-cues in the problem are congruent (nonconflict problems). This reflection is cognitive decoupling, especially indicative on poor performers who give a periodic correct answer (less than chance). The relationship between this index and conflict problem accuracy should be negative to support the TSM model prediction. However, the prediction was not supported: the cognitive decoupling index had a moderate positive correlation with conflict problem accuracy, $r(85) = .35$, $p = .001$, $R^2 = .12$. The intercept was not significantly greater than zero ($b = 467$ ms, $t(84) = 1.06$, $p = .29$). See Figure 3 for the pattern of responses on this index.

Individual differences analyses. Table 5 shows the correlations of performance on each problem type with the individual difference measures (CRT accuracy and NFC & AOT scale scores) and cognitive ability (SAT scores). There is a significant positive correlation of performance on the CRT with conflict problems only, indicating that people who did well on the CRT tended to do well on conflict problems (chose the base-rates more often). Additionally, performance on the CRT was positively related to a high Need for Cognition and cognitive flexibility (as measured by the AOT). Cognitive ability, as measured by self-reported SAT scores, was not associated with performance on the base-rate neglect task problems. A recent multiple mediation analysis (Swan & Revlin, 2015) was performed to test the relationship between behavioral inhibition (from the CRT) and how thinking disposition mediates onto a different task: conflict base-rate neglect problems. The analysis revealed that

although performance on the CRT was related to both the NFC and AOT, only the AOT was responsible for mediated performance on the conflict problems. The authors concluded that this inhibitive feature of dual process thinking is more related to the ability to be cognitively flexible and less related to domain-general motivational desires for higher cognitive activities. In this experiment, the results showed that answers on the CRT, a test that represents the ability (or inability) to inhibit an intuitive response, have a stronger positive relationship with performance on the base-rate neglect task better than general cognitive ability.

A recent Event Related Potential (ERP; De Neys et al., 2010) study investigated the electrical activity of bias-susceptible vs. bias-resistant reasoners on a dual process reasoning task. The authors found that the executive monitoring and inhibition functions were different for bias-susceptible reasoners (those who performed the "worst" on the task) when compared to bias-resistant reasoners (those who performed the "best" on the task). This study, among others, was the impetus to determine behavioral differences between participants who did well on the base-rate neglect task vs. those who did not do well. More specifically, with respect to the Predictions 2 – 4, is there a difference in RT between these two types of reasoners? In what ways can the data classify and identify the participants based on their responses?

To test the classification/ID hypotheses regarding performance, a median split was performed (*Mdn* = .40) on total conflict accuracy performance to answer these two questions and subjects were divided into bias-susceptible participants (*n* = 44) and bias-resistant

participants (n = 23).[5] A 2 x 2 x 2 mixed ANOVA was conducted to determine the extent of responses across problem types on RT by the two split groups. Table 6 contains the means and standard deviations for this test. Overall, there was a three-way interaction, $F(1, 65) = 4.82$, $p = .03$, $\eta^2_p = .07$. This suggested that the response patterns for the two groups were actually different, and it prompted further examination of the within-subjects effects. For bias-susceptible participants, a possible inhibition failure effect was observed (two-way interaction, $F(1, 43) = 30.38$, $p < .001$, $\eta^2_p = .41$): correct responses on conflict problems took significantly longer than incorrect responses, $t(46) = 4.09$, $p < .001$, $d = .54$. This is reversed for nonconflict problems, where incorrect responses are significantly slower than correct responses, $t(43) = 3.13$, $p = .003$, $d = .50$. For bias-resistant participants, there was no interaction between problem type and response ($F(1, 22) = .94$, $p = .34$): incorrect responses yielded slower RTs than correct responses, especially on conflict problems, $t(37) = 2.99$, $p = .005$, $d = .56$. The pattern was the same on nonconflict problems, $t(24) = 3.75$, $p = .001$. Bias-resistant participants are significantly slower when they choose a stereotype answer on either problem type (approximately three seconds slower on both).

A confirmatory discriminant analysis was performed to test the predictive power of the median split based on participants' accuracy performance on (a) both problem types, (b) RTs associated with correct and incorrect responses on both problem types, and (c) their CRT, NFC, and AOT scores. The 71 cases reported above were analyzed. A single discriminant function was calculated. The value of this function was significantly different for bias-susceptible and bias-resistant individuals ($\chi^2 = 86.90$, $df = 9$, $p < .001$). The

[5] Twenty-six participants were excluded from this analysis due to having no incorrect nonconflict problem judgments. Many of these participants had above-median accuracy.

correlations between the predictor variables and the discriminant function suggested that performance on conflict problems was the best predictor of bias (not surprisingly) on the task, followed by RT on these problems when the response was incorrect (stereotype). As expected, accuracy on conflict problems was positively correlated with the median-split groups, as higher values in the discriminant function would predict a bias-resistant individual; additionally, incorrect RTs on conflict problems had a similar positive correlation with the median-split groups, suggesting that bias-resistant individuals were more likely to spend more time on these problems. As evidenced by the correlations described above, the CRT, the NFC, and the AOT did not contribute to the discriminant function, suggesting that the bias classification is not located in thinking dispositions or cognitive style. Overall, the discriminant function successfully predicted the outcome for 98.5% of all cases, with accurate predictions made for 97.7% for the bias-susceptible group and 100% for the bias-resistant group (only one participant in the bias-susceptible group was predicted to be bias-resistant based on predictor variables).

Time series analyses. Figure 4 illustrates the average progression through all 50 trials for the participants. Overall, it appears that very little learning across trials is occurring, for either problem type. To wit, only nonconflict problems show a slight improvement across time ($R^2 = .10$, $p = .03$), whereas there was no improvement across time for conflict problems ($R^2 = .05$, $p = .13$). Furthermore, it appears that participants approached these problems individually without a sense that there was a pattern (likely due to the randomization), and even though the nonconflict trend over time was significant, the comparison of trends was not (Fisher's $z = .50$, $p = .61$).

The trial averages for each problem type were separated by the two groups described

above (bias-susceptible vs. bias-resistant). On conflict problems, there are two patterns to highlight: for bias-susceptible participants, they appeared to perform worse over time (R^2 = .09, p = .03); for bias-resistant reasoners, they appeared to get better over time (R^2 = .19, p = .001).

Discussion

The overall results of Experiment 1 indicate mixed support for the general predictions proffered by the LIM and TSM models of conflict detection and resolution. Behaviorally, most participants fell into the same trap set by Tversky and Kahneman (1974), utilizing the representativeness heuristic, mentioned earlier in the Introduction, to decide group membership rather than utilizing the basic base-rates of the problem. Notably this was more likely the case when the stereotypes and base-rates were incongruent. The general findings also support the idea that conflict detection is an essential part of the judgment process. Incorrect conflict judgments took longer than nonconflict judgments in general; this difference is a representation of internal, and likely implicit, conflict detection—but this is not the case for all participants (as shown in Figure 3). This finding provides support for the TSM, which states that conflict is imperfect and not universal (Pennycook et al., 2015), counter to De Neys' LIM (De Neys, 2012, 2014a).

With respect to the other behavioral RT-related hypotheses, support was mixed for inhibition (LIM) and decoupling (TSM). Overall group analyses pointed to inhibition: correct conflict judgments trended toward a longer RT, which would reflect an overall engagement of T2 to override the prepotent response of the stereotype. However, individual-level regressions showed that this relationship was negative, not positive, as individual accuracy increased, suggesting that the bias-susceptible individuals were quicker on correct judgments.

The bias-susceptible group showed that inhibition was a factor in correct responses. Bias-resistant participants did not reflect successful inhibition. Taken together, there was mixed support for the LIM based on group or individual analysis. Similarly, the cognitive decoupling hypothesis (TSM) was shaky: the direction of the relationship with correct conflict RT and nonconflict RT was opposite to predicted. It suggests that, again, the bias-resistant individuals spent more time decoupling rather than the bias-susceptible individuals. Overall, the behavioral findings suggest that T2 engagement occurred most often in the bias-resistant individuals. Finer-tuned ID analyses showed similar patterns.

The classification of individuals as bias-susceptible or bias-resistant was an attempt to classify individuals into multiple categories on all kinds of reasoning and judgment tasks. The pattern presented by the grouping of performance is an intriguing one. Bias-susceptible participants somewhat fit the hypothesized conflict detection and resolution behavioral pattern described by the TSM (Pennycook et al., 2015). When conflict is detected, it increases RT. The amount of added time is relatively small when compared to the amount of detection time given by bias-resistant participants, however. Moreover, in order for bias-susceptible participants to get the answer correct (which, on average, is not often), T2 is recruited, which performs the inhibition processing, leading to a base-rate answer ($r(45) = .24$, $p = .05$, one-tailed). This relationship is contrary to the finding that reflects the entire set of participants. It is additional evidence that bias-susceptible reasoners operate much differently than those who are relatively bias-resistant. Bias-resistant participants do not show this pattern of response or processing. They do show evidence of conflict detection, but do not show evidence of inhibition, since they tend to spend more time on problems that are ultimately stereotype. This finding is curious; is T2 recruited for rationalization, and not

decoupling? The overall findings promote this idea, since choosing the stereotype on any of these base-rate neglect problems would require rationalization for the incorrect choice and likely reflect deliberation and vacillation that operates against a defined base-rate strategy on these problems. As this is only one sample, subsequent experiments herein further investigated these results, with the opportunity to compare individuals across two types of reasoning/judgment tasks.

While the overall time series analysis did not show an overall learning effect, it did seem to show a practice effect. Figure 5 shows the average RTs over all trials. The larger RTs at the early stages represent a familiarizing with the task; this consequently reveals increased speed of response over time, while both problems types overlap and asymptote. Without unanalyzed practice problems to familiarize a person with the task prior to measuring decisions, RTs indicate that increased exposure to the problems quickened responses overall. When the accuracy effects are separated by group, some learning trends are discernible. The bias-resistant group tends to get better over time while the bias-susceptible group appears to get worse over time. This meaningful, because it suggests that in each group, an overall strategy was formed during the task, which led to base-rate decisions (bias-resistant group) or stereotype decisions (bias-susceptible group) more often as trials progressed. Experiment 3 will explore the strategy component in more depth when feedback of performance is manipulated.

While these data suggest that bias-susceptible and bias-resistant individuals approach the problems differently, there is not enough evidence to complete the overall description. The evidence does not support an inhibition failure bias from LIM, and since confidence/FOR was not measured on this task, it opens the possibility to tune the

classification of ID utilizing storage and monitoring/detection failures.

Furthermore, the base-rate neglect task is a social judgment task. Participants make inferences about people, not about objects or pretend situations. It is not necessarily a reasoning task *per se*; a test of the LIM/TSM and biases would be relegated to this task alone if no other reasoning is tested. In order for the LIM or TSM to be considered the cognitive framework for thinking in general, then it will need to explain and inform the observations on other judgment and reasoning tasks. Experiments 2 and 3 incorporate a conditional reasoning task to broaden the scope of the investigation.

Chapter III.

Experiment 2

This experiment continues the investigation into conflict detection and resolution, testing the LIM, the TSM, and the failure biases within them (De Neys, 2012, 2014a; De Neys & Bonnefon, 2013, Pennycook et al., 2015). The behavioral results of Experiment 1 show that there is mixed support for the LIM, but stronger support for the TSM. It appears that conflict detection is not a universal and routine process, inhibition does not necessarily account for correct answers for all individuals, and biased responding is marked by different behaviors depending on overall base-rate task performance. The results point to monitoring failures (failures of conflict detection) as the crucial issue for people who tend to choose the stereotypes more often on base-rate problems. The TSM aligns with this conclusion (Pennycook et al., 2015). This represents a lack of T2 engagement on the majority of the problems.

The goals of Experiment 2 were as follows: (1) To test the generality of the models described above by including an additional task within the reasoning realm: the LIM (De Neys, 2012; De Neys & Glumicic, 2008) and TSM (Pennycook et al., 2015) were developed on the base-rate neglect task; it is unclear if their specifications generalize to a reasoning task such as conditional reasoning (qualitatively distinct from a judgment task like base-rate neglect). Though conflict detection is the major component of these two models, if they were designed to describe and explain the framework of the cognitive architecture, the findings on a conditional reasoning task should be similar to and reflect findings from the base-rate neglect task (for a review of De Neys' boundary condition argument, see De Neys, 2014a). (2) To compare behavioral conflict monitoring with a metacognitive Feeling of Rightness

(FOR), which has been measured by confidence in one's answer, as a converging source of monitoring failures. Thompson and colleagues have argued that this metacognitive FOR signals T2 engagement. When confidence is low in an initial response to a problem or situation, deliberation may be required, and T2 is engaged. Conversely, if confidence is high in an initial response, T2 will likely be ignored for further deliberation (Thompson, 2009; Thompson et al., 2011; Thompson et al., 2013). This measure is inversely related to RTs, as low confidence would signal T2 engagement, thereby slowing responses; for high confidence, T1 responses will remain quick.

A similar methodology to Experiment 1 was used in the present study. However, several changes were made. First, the same base-rate problems from the first experiment were used. Second, a new, simple conditional reasoning task was added, utilizing some of the methodology of Thompson et al. (2011). The contrast in reasoning tasks can be compared directly, at the group and individual levels. In addition, performance replication on the base-rate task from Experiment 1 to 2 is a priority. Third, the AOT as an individual difference measure was dropped due to its length and its high correlation with the NFC scale. This was also a practical consideration, as the addition of the conditional reasoning task increased the length of the entire experimental session dramatically. Last, each question for both of the main tasks had a confidence rating scale after each problem.

Predictions

Research Questions 4 and 5 were addressed in this experiment. Below are the basic predictions:

Behavioral predictions. (1) Overall base-rate task performance will replicate Experiment 1. Participants will have lower accuracy scores on conflict problems than

nonconflict problems, while also responding slower on those conflict problems (correct > incorrect RTs) than nonconflict problems, on both types of reasoning tasks. (2) This pattern will additionally replicate on the conditional problems. (3) Confidence will be inversely related to response time, wherein low confidence will likely engage T2 thinking, increasing processing times and high confidence will not engage T2, decreasing processing times.

Individual differences hypotheses. In addition to testing the basic predictions of the metacognitive monitoring failure hypothesis, the LIM and TSM model predictions were tested once again on the (1) base-rate neglect task as well as (2) the conditional reasoning task. (3) The overall prediction for the comparison of failures/errors on the two tasks is that it will be task-dependent. The difficulty of the conditional reasoning task will be the likely cause of the dependency. For example, storage failures are a viable bias on the conditional reasoning task, because this kind of task requires knowledge of what validity means vs. whether a conclusion is believable or not, which is relatively automatic. Relatedly, storage vs. monitoring failures will not be behaviorally different.

Learning hypotheses. Over time, it was expected that Experiment 2 would follow the same patterns seen in Experiment 1. As each problem was randomized, it is likely that any given participant approaches each problem separately and without carryover from the previous problem. However, it was expected that overall confidence on both tasks would increase over time. This is due to the repetition factor—the more problems seen, the more likely a person becomes more confident in their answers.

Method

Participants. One hundred twelve undergraduates initially participated in this experiment for partial course credit. Ten participants were dropped from all analyses due to

incompletion of all experimental tasks (since most were within-subject variables). Thus, 102 participants (M_{age} = 18.92, *SD* = 1.12, 62% female) were included in overall data analysis. As with Experiment 1, if a participant did not contribute to the all cells in a given statistical test, they were excluded. Those sub-*N*s will be noted for each test.

Design and materials.

Base-rate task. This was the same as Experiment 1.

Conditional reasoning task. The conditional reasoning task was drawn heavily from Thompson et al. (2011) in order to test behavioral patterns from one reasoning/judgment task to another. A conditional reasoning task asks participants to complete an inference (drawing a conclusion) when the initial statement is presented in the form of "*if p, then q*". Four inferences can be made from this single form: Modus Ponens (MP), Modus Tollens (MT), Affirming the Consequent (AC), and Denying the Antecedent (DA). The first two inferences are logically valid (i.e., the conclusion follows necessarily from the premises), and the second two inferences are logically invalid. An example is as follows:

If a car runs out of gas, then it will stall.

The car has run out of gas. Therefore it will stall. (MP: valid)
The car has not stalled. Therefore it did not run out of gas. (MT: valid)
The car has stalled. Therefore it ran out of gas. (AC: invalid)
The car has not run out of gas. Therefore it will not stall. (DA: invalid)

There were 64 total problems, with 16 problems of each inference shown above. In order to create nonconflict and conflict problems, believability was manipulated. Half of the problems were believable, where *p* was a sufficient condition to bring about *q* in valid inferences or when *p* was a necessary condition to bring about *q* in invalid inferences; the other half of the problems were unbelievable, where *p* was not sufficient for *q* for valid inferences, or it was not at all necessary in the case of invalid inferences. Thus, there were

eight problems in each set of 16 that were believable and there were eight that were unbelievable. The problems either represented a causal conditional or a definitional conditional. All problems used for this task were originally developed in Thompson (1994). Refer to Appendix E for the entire set of problems.

For each conditional reasoning problem, the answer prompt was worded to mirror the base-rate task. For example, it appeared after the presentation of the two premises like this (from MP above): "*Is the conclusion likely to follow from the premises?*" This question was followed by a simple YES or NO response. A correct answer for valid conditionals was a YES response; for invalid conditionals it was a NO response. The results were coded as such for the dependent variable of accuracy.

Feeling of Rightness (FOR). Confidence ratings were added in this experiment as a measure of FOR. Participants responded to the question, "*At the time I provided my answer, I felt*:" They responded on Likert scale ranging from 1 to 7, where "1" represented a feeling of "*guessing*" and "7" represented a feeling of "*certainty*". The remaining label was given above the midpoint of the scale, representing a feeling of "*fairly certain*". This rating was given after each problem of both tasks and recorded a total of 114 times.

Thinking disposition and cognitive ability measures. This experiment used the CRT, NFC, and the SAT score from Experiment 1. The AOT was dropped due to the inclusion of the conditional reasoning task and its length vs. the length of the AOT and relative unimportance of the disposition measures in the previous experiment.

Procedure. Participants completed the each of the tasks at a computer station. Each student was introduced to the study and told that there were four stages to the experimental session. Participants first solved the three problems of the CRT. Participants then answered

the 50 base-rate problems. The order of these problems was fully randomized. Additionally, the answer was randomized, and it was either presented as the first option or second option (approximately 50% for each answer choice). Once the participant finished with those problems, they were given an opportunity to rest for 30 seconds. After the short rest period, participants answered the 64 conditional reasoning problems. The order of these questions was fully randomized. Once the participant finished these problems, they were given an additional opportunity to rest for 30 seconds. For all 114 problems across both tasks, participants were asked to rate their confidence in their answers after each individual problem, on the 1-7 Likert scale described above.

Upon finishing the two main tasks and the second rest period, instructions for the NFC appeared, describing the questionnaire and each of the corresponding scale values. The question order was randomized, and the scale appeared below each question. Finally, participants entered their most recent SAT score (out of 2400) and demographic information. Participants were debriefed, thanked, and dismissed. The entire experimental session lasted approximately 45-60 minutes.

Results: Base-rate Neglect Task

Accuracy. The base-rate task was analyzed in much the same way as in Experiment 1. First, to replicate the accuracy (choosing the base-rate answer for each problem), nonconflict vs. conflict problems were compared. The prediction was a replication of the robust representativeness bias on conflict problems. Table 7 shows the means (SDs) for accuracy and reveals that this prediction was supported: nonconflict problems had reliably higher base-rate selections than conflict problems, $t(101) = 19.20$, $p < .001$, $d = 1.86$.

Response time. The overall effect of RT was tested. Table 7 also shows that the

prediction that conflict problems would be slower than nonconflict problems (overall conflict detection) was supported: $t(101) = 8.88$, $p < .001$, $d = .41$. To test whether there was a distinct time difference between correct and incorrect responses on conflict problems, as well as nonconflict responses, an ANOVA was conducted. The difference between all three variables was significant, $F(2, 184) = 49.53$, $p < .001$, $\eta^2_p = .35$.[6] Moreover, the overall RTs for each in planned pair comparisons were reliably different (LSD, all pairs, $p < .001$). There was a large difference between correct conflict and incorrect conflict responses, with the former leading to three additional seconds of processing, on average. This large three-second processing gap supports the LIM and TSM models and their claims that T2 processing is the function of seconds-level processing/deliberation. This also offered support for conflict detection and inhibition failure on the group level, but more individual parsing is needed.

Conflict detection index. Experiment 1 offered mixed support for LIM model on the conflict detection and individual differences prediction. Figure 6 illustrates the positive relationship between the conflict detection index (incorrect conflict RT – nonconflict RT) and performance on conflict problems. This supported the TSM's position that conflict monitoring is not routine or universal. This prediction was tested on this second sample. At the group-level, as stated above, the two measures in the index had distinct time stamps ($t(92) = 4.62$, $p < .001$, $d = .33$). To reiterate, support for the LIM requires no correlation between RT difference scores on this index and conflict performance, which would indicate relatively stable detection for all participants. To replicate support for the TSM, a positive correlation for that relationship is required. A positive correlation was found: $r(91) = .39$, p

[6] This test included 93 participants. Nine participants were excluded because they did not make any errors on conflict problems.

< .001, R^2 = .15. Additionally, the intercept of this relationship was not significantly different from zero, which represents the lack of conflict detection or a mechanistic monitoring failure (b = -561 ms, t(91) = -1.09, p = .28) for a predicted individual who did not get a single problem correct on conflict problems. This relationship replicates the effect observed in Experiment 1.

Inhibition index. The inhibition index (conflict correct RT – conflict incorrect RT) was again tested in this experiment. This is based on the LIM's prediction of group differences applied on an individual differences level (De Neys, 2012). For the LIM, this index should increase as conflict accuracy increases, reflecting a greater reliance on inhibition for those individuals who choose the base-rate more often. For TSM, there is no direct model prediction (inhibition is deemed to be a function of cognitive decoupling). Figure 6 shows that there was a negative correlation: r(91) = -.30, p = .003, R^2 = .09. There was a positive intercept to the regression line, indicating that low accuracy individuals did inhibit somewhat to achieve some correct answers on conflict problems (b = 4.06 s, t(91) = 5.23, p < .001), but that those who chose the base-rate more often tended to be faster making correct responses than incorrect responses. This effect again replicated the observations of Experiment 1.

Cognitive decoupling index. Last, the TSM model prediction of cognitive decoupling was tested. The cognitive decoupling index was created by subtracting nonconflict RT from conflict correct RT. The support for the TSM would show a negative relationship between this RT difference and conflict problem accuracy. Figure 6 shows that there was no relationship between the two measures, r(91) = -.04, p = .70, R^2 = .002. The intercept was significantly different from zero (b = 3.50 s, t(91) = 5.22, p < .001), suggesting that correct

RTs are usually marked by positive differences with nonconflict RTs. This finding does not support the TSM and is not the same observation of Experiment 1.

Feeling of Rightness (FOR) and metacognitive monitoring. This is a similar concept to the LIM/TSM models organization of conflict detection; if a person has fluency of answer generation, then the answer comes to mind fast (T1) and yields high metacognitive confidence. If T1 generates multiple responses initially (LIM/TSM: "conflict"), then metacognitive monitoring should produce lower confidence if the conflict is detected. These hypotheses were tested in two ways: (1) group analysis with average responses and (2) individual-level analysis with correlations with behavioral conflict monitoring (conflict detection index). The latter should maintain a negative relationship: as confidence (FOR) increases, then the conflict detection index value (RT difference between conflict incorrect responses and nonconflict overall) should decrease. If a person does not detect a conflict, his/her conflict detection index value will be low, but there will likely be a high metacognitive FOR. A 2 (Problem: nonconflict and conflict) x 2 (Response: correct and incorrect) ANOVA was conducted to test the group data on FOR responses. While there was no main effect of Problem (overall nonconflict and conflict had similar ratings of FOR). Nonconflict problems had an average FOR rating of 4.07 (SD = 1.36) while conflict problems had a lower average rating of 3.74 (SD = 1.31). There was a Response main effect, $F(1, 71) = 17.44$, $p < .001$, $\eta^2_p = .20$, which led to a crossover interaction of the two variables: $F(1, 71) = 113.01$, $p < .001$, $\eta^2_p = .61$. On conflict problems, participants were more confident with incorrect answers than correct answers. On nonconflict problems, participants were more confident with correct answers and this difference is much more pronounced (see Figure 7, Panel A).

There were marginal, negative correlations between metacognitive FOR and the detection index. Nonconflict problems had a stronger correlation overall, $r(91) = -.17$, $p = .055$, one-tailed (marginal, right direction). A more conclusive relationship would be the relationship on conflict problems, but this relationship was weaker, $r(91) = -.15$, $p = .08$, one-tailed (marginal, right direction). This is mixed converging evidence. On the one hand, the direction of the effect is essential for this theoretical relationship, with metacognitive FOR, measured by confidence in an answer relating to the monitoring of a situation or problem and the subsequent engagement of a T2 processing. This inverse relationship carries over to the behavioral conflict detection measure directly based on RT. However, statistically, there is too much noise in one or both of the measures for the threshold of statistical significance.

An additional Analysis of Covariance (ANCOVA) was conducted with average conflict and nonconflict FOR entered as covariates into the 2 (Problem) x 2 (Response) model with RT as the dependent measure; this model was rather conclusive, as neither covariate was a significant predictor in the model and it did not appear that confidence has any moderating effect on RTs (adjusted means were identical to the unadjusted means in the 2 x 2 ANOVA).

Results: Conditional Reasoning Task

Analysis of the conditional reasoning proceeded identically to the base-rate neglect task. An effort will be made in this section to discuss results of the task and dependent measures within this task, as well as a comparison to the results of the base-rate neglect to track if patterns of behavior and cognitive processes extend to a more traditional reasoning task.

Preliminary analysis. To match the base-rate task and the dichotomy of nonconflict

and conflict problems in the base-rate neglect task, the conditional reasoning task was set up to match these conditions. The nonconflict set was comprised of valid-believable problems and invalid-unbelievable problems. The conflict set was comprised of valid-unbelievable and invalid-believable problems (see Evans et al., 1983, for a review). Before these problems were combined, comparison analyses were conducted to determine if a given response was more likely or faster than another (since accuracy is the method of performance, and not conclusion acceptance). On all problems, participants responded with equal speed to YES and NO responses across problems (ts(101) < 1.65, ps > .10). Additionally, a validity by belief ANOVA with accuracy as the dependent was conducted. Importantly for the combination of the above pairs, there was a significant logic and belief interaction (F(1, 101) = 297.66, p < .001, η^2_p = .75); accuracy for the nonconflict problem constituents was significantly greater than chance performance (.50; ps < .001), while the averages for the conflict problem constituents were significantly less than chance (ps < .02). These tests supported the combination of nonconflict and conflict problems for the remainder of analyses and for comparison with the effects of the base-rate neglect task.

Accuracy. After combining the problem types into nonconflict and conflict to match the base-rate neglect task, similar analyses were conducted to determine if effects were task-general. Table 7 displays the means (SDs) and reveals that was an accuracy effect, whereby participants were generally more accurate on nonconflict problems than conflict problems, t(101) = 17.25, p < .001, d = 1.97. Though the mean accuracy for nonconflict problems was lower on this task than the base-rate task, the pattern is similar for both problems, and as the reader can tell, the effect size is massive.

Response time. A one-way ANOVA was conducted to determine if there was general

conflict detection occurring, as well as overall T2 engagement. This test compared overall average nonconflict RT to average incorrect RT (conflict detection) and average correct RT (T2 engagement) on conflict problems. There was a significant effect: $F(2, 202) = 4.76$, $p = .01$, $\eta^2_p = .05$. See Table 7 for mean RT (SDs) on this task. Planned paired difference tests revealed that correct conflict solutions took reliably longer ($M = 8.8$ s, $SD = 3.46$, $p = .01$) than incorrect conflict solutions ($M = 7.9$ s), marking general T2 engagement to arrive at a correct answer (by definition, the *logical* answer). There was no difference between nonconflict solutions ($M = 8.1$ s) and incorrect conflict solutions, and the latter was overall faster than the former. This result reflects a lack of overall conflict detection by participants.

Conflict detection index. The conflict detection and monitoring hypotheses of the LIM and TSM were also tested. To reiterate the crucial predictions, the LIM postulates a flat relationship of the conflict detection index and a baseline of nonconflict RT (i.e., RT difference > 0). Alternatively, the TSM postulates a positive slope, with conflict detection increasing as accuracy on conflict problems increases. Figure 8 shows that this relationship was positive, but weak: $r(100) = .17$, $p = .04$, one-tailed, $R^2 = .03$. However, as evidenced by the above group test, the intercept was significantly below zero, meaning more often participants were faster to get a conflict problem wrong than make any response on a nonconflict problem ($b = -1.2$ s, $t(101) = -2.03$, $p = .05$. This test combined with the other suggests that the majority of participants were not detecting conflict on this task and were generally making many errors on these problems.

Inhibition index. While conflict correct responses were overall slower than conflict incorrect responses, what is the trend of inhibition for participants as a function of conflict performance (LIM)? There was no relationship: $r(100) = .06$, $p = .26$, one-tailed, $R^2 = .003$,

see Figure 8. Additionally, the intercept was not significantly different from zero, which means that some participants were faster on correct responses, while other were slower (b = 356 ms, $t(101) = .36$, $p = .72$. As with the conflict detection hypothesis, there isn't clear evidence of inhibition on conflict problems. Some participants did slow when making a correct conflict decision, but with the intercept marked at less than half a second with a relatively flat slope, it is difficult to suggest that the source of the errors are inhibition failures or monitoring failures.

Cognitive decoupling index. Do participants at least show a decoupling or rationalization pattern? The relationship between this index and conflict problem accuracy should be negative: less decoupling and therefore faster processing is needed as one makes more correct responses (TSM). The observed relationship in Figure 8 was weak, but in the opposite direction to the hypothesis, $r(100) = .16$, $p = .05$, one-tailed, $R^2 = .03$. However, the intercept was not significantly different from zero and negative: b = -828 ms, $t(101) = -.86$, $p = .39$. This is not surprising, since there were no reliable differences among the RT measures. Participants were no more likely to spend more time on correct responses (to decouple) than any effort placed on nonconflict problems. In many cases, responses were faster.

Feeling of Rightness (FOR) and metacognitive monitoring. A 2 (Problem) x 2 (Response) ANOVA was conducted with FOR as the dependent measure.[7] There was no main effect of Problem ($F(1, 99) = .38$, $p = .54$), nor a main effect of Response ($F(1, 99) = 2.41$, $p = .12$). In the case of two Problem types, the average FOR was equal for nonconflict and conflict responses. FOR was rated the same for correct and incorrect responses on the

[7] This test included 100 participants. Two participants were excluded for not having any incorrect nonconflict or conflict responses.

task. There was a significant crossover interaction: $F(1, 99) = 69.62$, $p < .001$, $\eta^2_p = .41$. This is because on nonconflict problems, responses have approximately a .60 mean difference on accuracy (correct: $M = 5.33$, $SD = 1.19$; incorrect: $M = 4.76$, $SD = 1.28$; $t(99) = 7.15$, $p < .001$, $d = .46$). This pattern is flipped for conflict problems, with a -.44 mean difference ($N = 102$; correct: $M = 4.86$, $SD = 1.28$; incorrect: $M = 5.30$, $SD = 1.17$; $t(101) = -6.70$, $p < .001$, $d = .36$). As Figure 7, Panel B illustrates, it appears that participants were more confident in their nonconflict correct answers than incorrect answers (as one would reasonably expect), but were overall less confident in their correct answers than their incorrect answers. On its face, this might make sense if there was some conflict detection occurring: T2 engagement for the correct answer would require a lower sense of FOR, which would in turn increase RT. The overall data do not support these abstractions, however.

The same 2 x 2 ANCOVA with RT as the dependent measure from the base-rate task was conducted with nonconflict and conflict FOR as covariates to determine if they account for any of the variance with RT. However, much like the base-rate task, FOR was not a significant predictor of RTs and it did not have any moderating effect on RTs from the original stats model (adjusted means were identical to unadjusted means).

Last, correlations were run between nonconflict FOR and the conflict detection index, as well as conflict FOR and conflict detection index. Both correlations were effectively zero (nonconflict: $r(100) = .01$, $p = .94$; conflict: $r(100) = -.02$, $p = .81$).

Results: Individual Differences and Time Series

Individual differences. Table 8 contains the correlations between the accuracy and FOR of both tasks and the cognitive ability (CRT and SAT) and thinking disposition (NFC) measures. There were a number of interesting relationships between the individual

differences measures and the base-rate and conditional reasoning performance/confidence measures.

CRT scores were positively correlated with both SAT and NFC scores, suggesting that ability measures converged and ability was related to a domain-general desire for complex cognitive activities. CRT scores were not correlated with accuracy or FOR ratings on the base-rate task, but there were small correlations between CRT scores and conditional reasoning conflict problem accuracy, as well as with both problem type FOR ratings.

In contrast with CRT, SAT scores were positively correlated with nonconflict and conflict accuracy on the base-rate task but not the conditional reasoning task. It was also positively correlated with the base-rate conflict detection index (not reported in Table 8, $r(82) = .22$, $p = .04$). This suggests that general cognitive ability is related to conflict monitoring.

Interestingly, NFC scores were only correlated with FOR ratings for both tasks. This likely reflects a motivational component described by the scale. If one desires to engage in complex cognitive activities, then confidence ratings on tasks designed to be complex cognitive activities should generally be positively related. However, the motivational component expressed here by the confidence ratings and NFC composite scores is beyond the scope of this investigation. Perhaps the more glaring result is that neither reasoning/judgments task was correlated with this thinking disposition scale, which is contrary to the results of Experiment 1 and other findings (e.g., Pennycook, Trippas, et al., 2014; Svedholm-Häkkinen, 2015).

Base-rate neglect. A similar median split analysis[8] from Experiment 1 was conducted on this dataset (*Mdn* = .40) and two groups were created: bias-susceptible participants (*n* = 48) and bias-resistant participants (*n* = 24). There was no three-way interaction between bias group, problem type, and response on RT or FOR ratings. Bias-susceptible participants were reliably faster overall than bias-resistant participants (approximately 2 seconds, $F(1, 70) = 4.55$, $p = .04$, $\eta^2_p = .06$; there were differences between the two groups on FOR ratings. The pattern of responses (large crossover interaction of problem by response) for bias-susceptible participants on RT and FOR replicate Experiment 1. In this experiment, bias-resistant participants also show this same crossover interaction, with slightly slower RTs overall. This effect is distinct from that observed in Experiment 1. Thus, according to this analysis, there were no individual differences between the two groups, contrary to the results obtained previously.

A confirmatory discriminant function analysis was also conducted. Utilizing the same predictor variables (minus the AOT), the function was able to differentiate the two groups ($\chi^2 = 82.94$, $df = 8$, $p < .001$). However, other than accuracy on the task, no other predictor could successfully place participants into susceptible or resistant groups.

Conditional reasoning. An additional median split analysis[9] was conducted on the new conditional reasoning task (*Mdn* = .38). There were no reliable differences between bias-susceptible participants (*n* = 52) and bias-resistant participants (*n* = 48) on RT or FOR ratings. Bias-susceptible participants trend faster overall responses than bias-resistant

[8] The test included 72 participants. Thirty participants were excluded for not contributing values for RT or FOR in each cell of the statistical model.

[9] The test included 100 participants. Two participants were excluded for not contributing values for RT or FOR in each cell of the statistical model.

participants, but this difference is approximately 400 ms. This is not enough RT difference to differentiate response patterns and it does not reflect additional processing beyond T1 (which bias-resistant participants who detect conflict would move beyond T1 into T2).

The discriminant function analysis revealed a lack of clear distinction between bias-susceptible and bias-resistant groups using the same predictors as the base-rate neglect task ($\chi^2 = 14.86$, $df = 8$, $p = .06$). This finding represents additional clear evidence that the two tasks are distinct.

Time series analysis. Like Experiment 1, tracking how participants did over time can help explain how participants complete these tasks. The same analyses as in Experiment 1 were conducted here.

Base-rate neglect. There was a large difference in accuracy across all problems. Though both regression lines were positive, there were no reliable practice or learning effects across the set of problems averaged over participants. For RT, practice effects were observed on both problem types. Conflict problems overall were slower than nonconflict problems, but the relationships were not significantly different from each other. This was likely due to the first few problems causing participants to take a little more time before answering since they likely had not seen the task before. Once they were acclimated to the task, they began to quicken. In the case of the FOR measure, average confidence rested between points 4 and 5 on the scale across problems and participants (which is above the midpoint of the scale). Conflict problems engendered a slightly lower confidence overall, but the trends were effectively parallel.

The accuracy and RT effects over trials replicated Experiment 1. The trends in this experiment were slightly flatter for accuracy than Experiment 1; the RT trends continued to

reflect practice effects.

Conditional reasoning. Participants tended to get less accurate over time on nonconflict problems (r = -.39); conversely, they appeared to get more accurate on conflict problems (r = .27) over time. Noticeably, nonconflict problems got much more variable and erratic as more problems were tackled. As evidenced by the group tests discussed above, RT differences were essentially nonexistent in this data set. RT time series revealed that nonconflict and conflict problems had the same pattern. Participants started slow and sped up as they went through the problems, showing a practice effect on the first few problems due to novelty. In the case of the FOR measure, flat relationships were observed. Confidence ratings were above 5 on the scale and do not vary much from there across all problems.

Since this task was not in Experiment 1, no cross-experiment comparison can be made. However, these patterns can be compared to the base-rate neglect task. The RT and FOR measures are essentially the same for both tasks: RT trends reflected a practice effect and FOR trends were as flat as the base-rate task. Accuracy trends deviated in this task compared to the base-rate task.

Discussion

The goals of Experiment 2 were to (1) extend two existing dual process models in base-rate neglect to a qualitatively different reasoning task, that is, conditional reasoning (Markovits, Thompson, & Brisson, 2015; Thompson et al., 2011; Thompson, 1994). Additionally, the models (LIM and TSM; De Neys, 2012; Pennycook et al., 2015) tested in Experiment 1 were extended to this dataset for replication and extension purposes, and (2) explore the notion of metacognitive monitoring within the DPTs framework utilizing a measure of confidence in one's answer as evidence for behavioral decisions (e.g., Thompson

et al., 2011).

Base-rate neglect. On the base-rate task, the general findings of Experiment were replicated. There was a robust conflict detection effect that was replicated. Accuracy effects were replicated, as well. There was a clear distinction between nonconflict preferences for the base-rate answer (high) vs. this preference on conflict problems (low). However, stronger effects were observed here for some hypotheses. For example, there was support for behavioral inhibition measuring RTs at the group level (De Neys & Glumicic, 2008), which was not observed in the previous experiment. The conflict detection index effect was reduced, but there remained a significant relationship between conflict detection and choosing the base-rate answer on conflict problems. Though the group test for inhibition yields support for the LIM (De Neys, 2012; De Neys, 2014a), the inhibition index is not supportive of this model. A negative relationship was observed between inhibition RT difference scores and the choice of the base-rate answer on conflict problems, suggesting that as accuracy increases, participants get faster on correct answers to the task. Finally, there was no support for the cognitive decoupling index's relationship with base-rate answers on conflict problems.

On the face of these observations, there is still mixed support for both the LIM and TSM, with the strength of evidence tilted toward the conclusion that conflict detection is not routine and universal (TSM; Pennycook et al., 2012; Pennycook et al., 2015). Inhibition evidence is contradictory to its conceptual definition—bias-resistant participants are not engaging in inhibition (T2) to achieve correct answers (De Neys, 2012). This finding, along with that of Experiment 1, are not aligned with the idea in LIM (and even the TSM) that T2 engagement always follows from successful conflict detection. In actuality, the bias-resistant

participants in this experiment rarely spent much extra time considering the base-rate after realizing it was a better answer than the stereotype. It appears they actually utilize T1 processing whereby the logical intuition is the more salient response, leading to an overall faster answer (Handley & Trippas, 2015; Pennycook et al., 2015). The cognitive decoupling index supports this idea: decoupling reflects the time difference between inhibition and detection, and these two indices are in direct opposition with each other. The final piece of this puzzle will be explored in Experiment 3 with an additional dataset to explore.

The second important piece to this effort was the role of metacognitive monitoring (Thompson, 2009; Thompson et al., 2011) in the form of a Feeling of Rightness (FOR). A high FOR is marked by high confidence in an initial response and fluency of response generation (Thompson et al., 2013). Overall, it was hypothesized that a failure of metacognitive monitoring is a failure of conflict monitoring, which is in turn a failure of conflict detection (De Neys & Bonnefon, 2013; Thompson et al., 2011). On the base-rate neglect task, FOR differences were observed as an interaction of problem type and response on those problems. Higher FOR ratings were generated more often when a person made a stereotype choice on conflict problems than when a base-rate choice was made. This pattern reverses on nonconflict problems. The ratings of observed confidence and the differences between groups match the outcomes of Thompson et al.'s (2011) Experiments 1 and 2, in which confidence differences were small but invariable within groups. These effects point to stability participants expressed throughout the task. It makes sense: the problems do not change in structure, merely content. Regardless of this stability, however, the full interaction of problem and response reflects overconfidence, a reflection of the saliency of the stereotypes on the conflict problems (Pennycook, Trippas, et al., 2014; Swan & Revlin, 2015;

Svedholm & Lindeman, 2013), rather than the making uneasy base-rate selections. Crucially, these FOR ratings did not interact with RTs, suggesting a possible hindsight bias with the confidence rating appearing after the problem and decision.

The marginal correlations of FOR and the conflict detection index yet still offer mixed support as converging evidence that metacognitive FOR reflects a general conflict monitoring failure. More data is needed and additional analysis will occur in the next experiment.

Conditional reasoning. The comparison of effects from the base-rate task and conditional reasoning tasks was empirically and theoretically important to determine if the models (LIM and TSM) created based on base-rate neglect transferred to other tasks (task-general or task-specific models). First, the base-rate neglect task relies on the heavy-handed distinction between probability (arguably an implicit intuition among most adults; Pennycook & Thompson, 2012; Pennycook, Trippas, et al., 2014) and person categorization/relative ease and adaptive quality of stereotypes (Hutchison & Martin, 2015; Kahneman & Tversky, 1973; Tversky & Kahneman, 1974). The representativeness heuristic within the base-rate task is robust and is rarely thwarted (regardless if you ask for probability judgments, e.g., in Kahneman & Tversky, 1973 or Thompson et al., 2011, or if you ask for a dichotomous forced choice, e.g., in De Neys & Glumicic, 2008 or Pennycook et al., 2011). On the other hand, in conditional reasoning, you have a class of tasks that relies heavily on logical deductive reasoning (e.g., syllogistic reasoning or conditional reasoning), wherein the structure of the premises and conclusions is the sole contributor to validity. These tasks add conflict by manipulating the believability of the premises, conclusions, or both, by modification of the semantics of the statements. Invariably, there are more moving parts to

this type of reasoning task than the base-rate neglect task. The purpose of comparing these two distinct tasks is not to state which task is better or better reflects human intelligence or rationality, but to determine whether the current dual process models (namely TSM and LIM), developed with the base-rate neglect task, can describe and explain behaviors on a qualitatively different task. If they cannot, a new model that task-general is theoretically warranted.

Creating nonconflict and conflict problems to match the structure of the base-rate task yielded a similar accuracy effect: nonconflict problems had significantly more correct answers (choosing YES when an argument was valid and choosing NO when an argument was invalid) than conflict problems. However, the replication of effects ends here. RT analyses revealed a lack of conflict detection by group or individual-level analyses; many participants were actually faster to get a conflict problem incorrect than make any decision on nonconflict problems. There was an inhibition effect (LIM), but only at the group level. Variation among RTs on the conditional reasoning was too similar, reflecting very limited T2 activation among all participants (bias-susceptible and bias-resistant). The conflict detection, inhibition, and cognitive decoupling indices were either not different from zero or in the wrong direction, erasing all support for the model predictions of both the LIM and TSM.

Perhaps the only meaningful comparative finding between tasks was an interaction of problem and response on FOR. The same overconfidence effect was observed on conflict problems. Thus, a high sense of confidence in an answer reduces T2 engagement, leading to a frequent incorrect solution. The link between FOR and T2 engagement is replicated across tasks, but again, RTs to accurately reflect T2 engagement were not affected by FOR ratings. In addition, there was no relationship between FOR ratings and the conflict detection index.

This is likely due to the lack of conflict detection on this task in general; FOR appears to be the more sensitive measure of conflict monitoring than RT on this task; this also aligns with the results of Thompson et al. (2011).

In sum, the results of Experiment 2 offered support for the idea that conflict detection is imperfect, relating to the idea that monitoring failures are an important individual difference to consider in this dual process framework (Handley & Trippas, 2015; Pennycook et al., 2012; Pennycook et al., 2015). The results supported the model predictions of the TSM over the LIM. The base-rate task is an important piece to these two dual process models, but they are likely attributable in this way because of their development with the task (De Neys, 2012, 2014a; De Neys & Glumicic, 2008; Pennycook et al., 2015). Conditional reasoning does not translate to the model predictions of the TSM and the LIM; these models are task-specific. These tasks were re-used in Experiment 3 to continue to collect more data for replication and extension purposes. Finally, are the individual differences failure accounts (monitoring, inhibition) modular (De Neys, 2012, 2014a; De Neys & Bonnefon, 2013) and would a manipulation of feedback interact with conflict monitoring and FOR? Experiment 3 utilizes a feedback manipulation to address this question.

Chapter IV.

Experiment 3

The first two experiments of this dissertation attempted to address the two recent models in DPTs, integrate more individual-level analyses beyond that of group effects (De Neys, 2014a; Pennycook et al., 2015), and apply a task-general lens to these two models that have been developed solely with base-rate neglect. Experiment 3 adds to these endeavors, but focuses more on the role of the role of individual differences (storage, monitoring, and inhibition) and whether they are actual modal processes, as argued by De Neys and colleagues (De Neys, 2014a; De Neys & Bonnefon, 2013). The substantial way to test this postulate is to offer feedback to reasoners within both tasks that is meant to interact with conflict monitoring and detection. Thus, through this experiment, I have linked the model predictions of the LIM and the TSM, the hypotheses of monitoring failures as failures of conflict detection, and that conflict monitoring is an associated concept to metacognitive monitoring, as reflected by FOR.

The present experiment is designed to offer feedback that would interact mainly with confidence in an initial T1 response. Without any feedback, individuals solving reasoning problems or making judgments would likely utilize a strategy they find most salient. Bias-susceptible people would rely on initial T1 responses when it is strongest or salient, whereas bias-resistant people would typically utilize better strategies for more normative responding. In both cases, an initial FOR would likely dictate the course of responding—high FOR, less T2 engagement; low FOR, more T2 engagement. The type of feedback given here is important because it needs to impact confidence/FOR in a meaningful way. Other than explicitly telling a person that they are approaching a given task incorrectly, confidence

could be impacted by feedback that is designed to reflect the strength of performance (a percentage correct or related measure) or relative standing with other who have completed the task in the past. Manipulating this feedback can have the desired impact on confidence without directly addressing the normative strategies necessary to complete a base-rate neglect task or a conditional reasoning task. Previous research has shown that offering feedback can shift a decision criterion in a recognition task (Aminoff et al., 2012), increase counterfactual thinking (a T2 function; Roese, 1997), and induce positive or negative emotions depending if false feedback was given as excellent or poor, respectively (Evers, de Ridder, & Adriaanse, 2009). A central question is can this feedback interact enough with the initial FOR to induce a change in behavior that is indicative of T2 engagement, such as increased processing times after feedback is given (Thompson et al., 2011)? Would this effect reflect the modal nature of the failure hypotheses (namely monitoring)?

In DPTs and reasoning research described above, few studies (e.g., De Neys, 2014b) have offered the participant feedback for their performance on individual problems. In the studies conducted by De Neys, utilizing the LIM as the basis of his predictions, he found that participants were faster and more accurate on conflict problems when given feedback. He argued that the faster RTs on conflict problems reflect an interaction with inhibition failures—participants had to have already recognized the conflict (conflict detection). Feedback did not interact with monitoring failures: it did not help people who did not initially realize there was a conflict between the heuristically-cued response and the normative/probabilistic/logically-cue response (De Neys, 2014b).

The LIM predicts and is supported by these data from De Neys (2014b). Additionally, he conducted the studies utilizing the conjunction fallacy task (the “Linda” problem;

Kahneman & Tversky, 1973) and categorical syllogisms (De Neys, 2006). The latter is important for my studies—it appears he garnered support for his model that spans two qualitatively different tasks (social inference and logical inference). However, the feedback manipulation used in De Neys (2014b) was not intended to interact with confidence in responses, which is a possible contributor to monitoring issues not addressed in the LIM.

Experiment 3 was designed to manipulate a participant's confidence (FOR) in their answers. Specifically, falsely suggesting a participant's performance is worse than a fictitious dataset or giving them true feedback was intended to decrease confidence in previous decisions on a set amount of trials. In contrast, falsely suggesting a participant's performance is better than a fictitious dataset was intended to inflate confidence and possibly increase erroneous responses post-feedback. All participants solved the same 50 problems on the base-rate neglect task and conditional reasoning task from Experiments 1 and 2. Halfway through each task, a rest break was initiated and the participant was presented with information regarding their performance. The False-High feedback group received two pieces of information regarding performance: a predetermined high percentage correct and the relative standing that their performance was better than the top 10% of people who have completed the same task. The False-Low group was the opposite: they received a predetermined low percentage correct and given the relative standing that their performance was worse than the bottom 10% of people who have completed the same task. The True feedback group received their actual, calculated percentage correct without information regarding their relative standing. The Control group did not receive feedback (they would still receive a break halfway through the trials). This final control group was utilized for replication purposes to compare behavioral outcomes with Experiments 1 and 2, as well as to

act as a no-manipulation control for the feedback manipulation variable.

Predictions

The question of whether providing feedback on these two tasks would interact with confidence (Research Question 6) was addressed in this experiment. Basic predictions were as follows for this experiment: (1) The False-Low and True feedback groups will have better accuracy on conflict problems than the False-High or Control groups post-feedback because the instigation of lowered confidence should decrease FOR and increase T2 engagement for more normative/logical responses. (2) Similarly, RTs overall and on conflict problems will be higher for the False-Low and True feedback groups post feedback than the False-High or Control groups, which reflects an increase in conflict monitoring efficacy and T2 engagement (Thompson et al., 2011). (3) There should be an interaction of feedback group and feedback presence (pre- or post-feedback) on accuracy, RT, and FOR. (4) The previous predictions tested in Experiments 1 and 2 were also tested here within the feedback manipulation.

Method

Participants. One hundred thirty-five undergraduates participated in this experiment for partial course credit. As with Experiment 2, participants who did not complete the task were excluded from all analyses (five participants were excluded from all analyses for not completing the task or leaving unrealistic responses on the base-rate or conditional reasoning tasks). Thus, 130 total participants were analyzed in this study (M_{age} = 19.2 years, SD = 2.59, 78% female). Participants who did not contribute to all variables in a statistical model were excluded, as per the previous two experiments. The sub-Ns are noted for each test.

Design and materials. A 4 x 2 x 2 mixed factorial design was utilized for this

experiment. The first independent variable was feedback group, a between-subjects variable. One group received feedback that was designed to increase their sense of confidence in their ongoing performance by suggesting that their performance was significantly *better on the task than the top 10% of people who have completed the task* (False-High group, $n = 36$). A second group received feedback that was designed to decrease their sense of confidence by suggesting that their performance was significantly *worse than the bottom 10% of people who have completed the task* (False-Low group, $n = 31$). A third group received minimal, but true, feedback—a direct computation of their performance through the first part of each task (True feedback group, $n = 34$). The feedback for these three groups was presented halfway through each task. Specifically, feedback was presented after the 25th problem on the base-rate task and after the 32nd problem on the conditional reasoning task. The three feedback groups each received a percentage of performance on the first half; the false groups received a fixed value that was either high or low, to match with the phrase of relative performance. This percentage was different for each task, so the message wasn't exactly the same from base-rate task to conditional reasoning task. The True feedback group saw only the percentage and no relative standing. The final group did not receive any feedback (Control group, $n = 29$), but did receive a break of equal length halfway through each task. The second independent variable (within-subjects) was the comparison between performance and decision of the first half (before feedback) and the second half (after feedback) of each task. The final independent variable (within-subjects) was the type of problem, described in the previous experiments: nonconflict problems and conflict problems.

Measures of interest matched those collected in Experiment 2: accuracy, response times (RTs) and confidence/FOR.

Base-rate task. This task was replicated from Experiments 1 and 2.

Conditional reasoning task. This task was replicated from Experiment 2.

Individual differences measures. These measures remained the same from Experiment 2.

Procedure. The procedure for this experiment matched closely with Experiment 2. The task order remained the same from the previous experiment, and question order (conflict vs. nonconflict problems) remained randomized. Changes that occurred for this study: A break after the 25th trial (on the base-rate task) and the 32nd trial (on the conditional reasoning task) was added for each group to display feedback. For the false feedback groups, a faux calculation score was displayed and underneath this score on the screen appeared the relative standing phrase according to the feedback group (above top 10% or below bottom 10%). The true feedback group simply saw a percentage score of their performance on the first half of the trials for each task. The control group only received a message to rest. The entire rest period for all four groups was 30 seconds. After the break, the remaining half of the trials per task began immediately. Following the NFC and demographic questions, two additional open-ended questions were asked. The first question was asked to all of the groups and specifically asked the participants to describe their strategy on the task. The second task was only presented to the feedback groups and it asked whether the participant changed their strategy after the feedback display. After completing the final strategy questions, participants were debriefed, thanked, and dismissed.

Results: Base-rate Neglect Task

Preliminary analyses. Analyses were conducted on the Control group to test whether effects replicated from the previous experiments. In the Control group, base-rate answers

were chosen more often on nonconflict problems than conflict problems, $t(28) = 8.74$, $p < .001$, $d = 1.69$. The overall RT effect replicated as well: nonconflict problems had the fastest responses overall with a significant increase to incorrect conflict responses, $F(2, 54) = 8.74$, $p = .001$, $\eta^2_p = .24$.[10] The planned paired comparisons revealed that nonconflict RT to incorrect conflict RT increase was significant (Mean Difference = 2.80 seconds, $p < .001$), but that the difference between incorrect and correct responses was not significant (Mean Difference = 1.35 seconds, $p = .29$), but in the predicted direction (a greater difference for these two levels than in Experiment 1). On FOR, nonconflict problems had a higher confidence rating, $t(28) = 5.00$, $p < .001$, $d = .42$. With these basic results essentially the same behavioral observations as the previous two experiments, the remaining hypothesis analyses were tested.

Second, the performance pre-feedback (Pre-FB) needed a confirmatory test to determine if all four groups went through the first half of the task similarly before any manipulation occurred (a pretest). The tests conducted were the same as the replication analyses. A 4 (Feedback Group) x 2 (Problem) ANOVA was conducted on accuracy. There were no Group differences ($F(3, 126) = .28$, $p = .84$) nor an interaction ($F(3, 126) = .28$, $p = .84$). There was an expected main effect of Problem, wherein all groups had higher accuracy on nonconflict problems than conflict problems, $F(1, 126) = 367.48$, $p < .001$, $\eta^2_p = .75$. The same 4 x 2 ANOVA was run on RTs; there was a marginal effect of Group, $F(3, 126) = 2.65$, $p = .051$, $\eta^2_p = .06$. It appears the False-Low feedback group was overall slower (approximately two seconds) than the other three groups before any feedback was given.

[10] For the RT analysis of the control group, one person was omitted from the analysis for not having any incorrect conflict responses ($n = 28$).

There was also an expected effect of Problem, $F(1, 126) = 63.66$, $p < .001$, $\eta^2_p = .34$. Though there was a minor difference of the feedback groups, there was no interaction of these two variables ($F(3, 126) = 2.46$, $p = .07$). On FOR, there was no effect of Group, $F(3, 126) = .40$, $p = .75$. There was only the expected effect of Problem, $F(1, 126) = 68.24$, $p < .001$, $\eta^2_p = .35$, whereby nonconflict problems were rated with higher confidence than conflict problems. No interaction of the two variables was observed, $F(3, 126) = .29$, $p = .83$. The pre-FB scores were then entered into post-FB statistical models as a covariate to account for performance before feedback manipulation. Special attention was paid to the False-Low feedback group to determine if the group was inherently different and confounds in selection existed.

Main analyses. The remaining analyses were conducted to test the main study hypotheses for the base-rate neglect task. Three main 4 (Group) x 2 (Problem) x 2 (Pre/Post-Feedback) ANCOVAs were conducted with accuracy, RT, and FOR as separate dependent measures. Pre-FB scores were entered into these statistical models as covariates to control for performance on these measures before the main feedback manipulation. Table 9 shows the means and standard deviations for the various effects described below.

Accuracy. First, there was no Group main effect; each group has the same relative average performance across problem type and feedback, $F(3, 125) = .26$, $p = .85$. Across the entire task, base-rates were chosen more often on nonconflict problems than conflict problems, as expected, $F(1, 125) = 367.30$, $p < .001$, $\eta^2_p = .75$. There was also a main effect of Feedback, whereby all the groups did slightly better after feedback than prior to feedback. This led to a Problem by Feedback interaction, because the increase was greater on conflict problems post-FB, $F(1, 125) = 7.04$, $p = .009$, $\eta^2_p = .05$. No other effects or interactions were

significant. Overall pre-FB accuracy was a significant predictor of accuracy, $F(1, 125) = 598.02$, $p < .001$, $\eta^2_p = .83$. This covariate interacted with Problem and Feedback variables (separately), Fs$(1, 125) > 8.64$, ps $< .004$, η^2_ps $= .07$. It did not interact with Group, however. This is likely due to the fact that with the covariate entered into the model, each condition improved after FB was given. The False-Low group appeared to have improved the most on conflict problems; however, this is increase is not significant. Interestingly, the control group got slightly worse on nonconflict problems post-FB. With minimal information given to help improve their problems, True feedback participants had the flattest increase on conflict problems.

Response time. There were limited effects on RT in the omnibus test. First, there was no Group main effect, $F(3, 125) = .58$, $p = .63$. Additionally, there was no overall Problem main effect, $F(1, 125) = .79$, $p = .38$. In addition, there was no main effect of feedback, $F(1, 125) = 1.33$, $p = .25$. No main interactions were significant (two-way or the three-way). The majority of the variance in this model was accounted by the pre-FB covariate, ($F(1, 125) = 1330.38$, $p < .001$, $\eta^2_p = .91$. This led to an interaction of the covariate and Problem, and the covariate and Feedback condition. The data show that everyone in all four groups got faster in the second half of the problems from the first half. In addition, the difference between conflict problems and nonconflict problems was approximately three seconds slower in all groups. Feedback did not differentially affect any of the feedback groups as predicted (e.g., a less steep RT difference of speed in the False-Low feedback group signaling a noticeable slowing of decisions).

FOR. A similar story to RT is present for confidence/FOR scores. First, there was no main effect present for Group, $F(3, 125) = .89$, $p = .45$. There was a main effect of Problem,

$F(1, 125) = 6.66$, $p = .01$, $\eta^2_p = .05$, but no main effect of feedback, $F(1, 125) = .05$, $p = .83$. However, this did not lead to a Problem by Feedback interaction, as predicted. Group did not interact with either variable. Pre-FB FOR ratings were a significant covariate in the model ($F(1, 125) = 876.96$, $p < .001$, $\eta^2_p = .88$), accounting for approximately 88% of the unique variance of FOR ratings. From all appearances, the manipulation *did* have the intended effect by reducing confidence ratings in the False-Low feedback group, but these mean differences are not significant—though they were similarly small when compared to the differences observed in FOR throughout this study. Interestingly, the feedback given in the False-Low condition actually decreased FOR ratings for nonconflict problems at a greater rate than conflict problems (adjusted means, Mean Differences = .49 and .31, respectively). Moreover, the confidence decrease in this group was significant on those nonconflict problems, which was not observed in the other three groups (see Table 9).

A final ANCOVA of Group by Problem on RT that included post-FOR as a covariate was conducted to determine if the reduction of confidence in the False-Low group lead to longer processing times (pre-FB RT was also entered into the model as a second covariate to match the test in the previous paragraph). Post-FB FOR was not a significant predictor of RT across problems. Adjusted RT means actually show that on conflict problems, the False-Low feedback group was the fastest of the four groups (not a reliable difference, however).

Conflict detection hypothesis. The relationship between conflict problem accuracy and the RT difference between incorrect conflict responses and nonconflict responses (conflict detection index) was explored across feedback groups and separately within the feedback groups. Across all participants in the study, there was clear support for the TSM, $r(119) = .50$, $p < .001$, $R^2 = .25$. Figure 9 illustrates the conflict detection index by group. It

shows that all feedback groups and the control group have the same overall relationship: *r*s > .40, *p*s < .05. None of the correlation coefficients are different from the others. Thus, in all groups, participants who selected the stereotype more often on conflict problems did not do well at detecting conflict, whereas participants who selected the base-rates more often on conflict problems were able to detect the conflict and continue to increase the time spent deliberating even when they got the problem incorrect. Additionally, the intercept was not significantly different than zero, or no RT difference, showing the lack of conflict detection for those participants who made more stereotype selections on conflict problems (-913 ms, $t(119) = -1.78$, $p = .08$).

Inhibition hypothesis. The relationship between the inhibition index (correct conflict responses – incorrect conflict responses) and conflict problem accuracy was tested. Low accuracy participants in all groups spent more time inhibiting on correct responses, but as performance on conflict problems increased, participants who made fewer errors spent less time to making a correct response than making an incorrect response in general (*r*s < -.36, *p*s < .05; across groups $R^2 = .27$, see Figure 9). The relationship is further support against the LIM. The overall intercept here was significantly greater than zero, representing the greater amount of time needed for inhibition for participants who made many errors (4.9 s, $t(119) = 6.58$, $p < .001$).

Cognitive decoupling hypothesis. For the entire sample, there was a negative relationship between the cognitive decoupling index and conflict problem accuracy, $r(118) = -.22$, $p = .02$, $R^2 = .05$. The intercept for this relationship is significantly greater than zero as well, 4.0 s, $t(119) = 5.90$, $p < .001$. Participants who made fewer errors spent more time on correct conflict judgments than overall nonconflict judgments in general. So participants who

made more errors, on the few correct conflict judgments they had, ultimately spent less time decoupling. This is the first time in the data presented that the TSM Decoupling hypothesis was supported. Additionally, each Group separately had a negative relationship between the two variables, but these were flatter relationships (*p*s > .06, see Figure 9).

Feedback interaction with detection, inhibition, and cognitive decoupling. Table 10 shows the correlations of the three above indices with conflict problem performance before and after feedback. The majority of the correlations are not significant (flat), but the presence of feedback seems to change behavior patterns for some participants. There were no observed correlation differences between pre-FB and post-FB.

Conflict detection and FOR. There was no relationship between conflict detection/monitoring and metacognitive monitoring (FOR), $r(118) = .03$, $p = .74$. Interestingly, parsing out the correlations by group reveals interacting relationships. The Control group and the False-Low feedback group each had negative correlations between the two monitoring measures. The other two groups, True feedback and False-High feedback groups each had positive correlations. However, with each groups relatively small sample sizes, these effect sizes were not reliably different from zero.

Results: Conditional Reasoning Task

Preliminary analyses. The same procedure used in the Experiment 2 analysis to combine the conditional reasoning task variables was utilized here. Overall, YES and NO responses across all problems were equally fast (*t*s(129) < 1.21, *p*s > .23). The validity by belief ANOVA with accuracy as the dependent variable revealed the important interaction between these two variables, $F(1, 129) = 554.70$, $p < .001$, $\eta^2_p = .81$. Correct decisions on valid-believable and invalid-unbelievable problems were significantly greater than chance

(ps < .001); invalid-unbelievable decisions were significantly less than chance (p < .001); valid-unbelievable decisions were not different from chance, but approximately 57% of participants' decisions fell below .50, which made a reasonable situation to combine these latter two problem types into conflict problems.

The same two preliminary tests from the base-rate neglect task were conducted on the conditional reasoning problems: First, the validity by believability ANOVA from the previous paragraph was replicated within the Control group; there was an interaction of logic and belief. The pattern of accuracy closely aligned with the overall dataset, but again, participants were close to chance on valid-unbelievable problems. The assumption to combine them with invalid-believable problems to create conflict problems was retained, however.

For overall accuracy, there was a large effect on accuracy between the problems, $t(28) = 10.87$, $p < .001$, $d = 2.47$. Participants in the Control group made correct decisions on nonconflict problems ($M = .78$, $SD = .12$) at a greater rate than conflict problems ($M = .33$, $SD = .15$). For the overall RT test, there was no effect between nonconflict RTs and conflict RTs, $F(2, 56) = 1.39$, $p = .26$. Planned paired comparisons confirmed that RT differences between nonconflict problems and incorrect conflict decisions were approximately 450 ms, while RT differences between incorrect and correct conflict decisions were approximately 380 ms. These values suggest that the Control group had trouble detecting the conflict on these problems, as well as that any additional time spent processing correct responses did not enter the realm of inhibition or decoupling. Though the accuracy test replicates Experiment 2's results, the overall RT test did not.

Last, nonconflict and conflict problems were compared on complete-task FOR. There

was no difference between overall FOR, with nonconflict problems having only a slightly higher FOR (M = 5.40, SD = .98) than conflict problems (M = 5.33, SD = .97), $t(28) = 1.40$, $p = .17$. I will conduct the remainder of the statistical tests, but any conclusions drawn from this sub-sample will need to be weighed against conflicting findings.

Next, pre-FB scores were tested on the three main DVs using the 4 (Group) x 2 (Problem) ANOVAs to determine if group differences on the conditional reasoning task existed prior to any feedback manipulation. There was no difference between the groups on accuracy, $F(3, 126) = 2.50$, $p = .06$. There was a curious Problem effect, whereby each group had better accuracy on conflict problems (M = .58, SD = .15) than nonconflict problems (M = .52, SD = .13), $F(1, 126) = 14.72$, $p < .001$, $\eta^2_p = .11$. There was no interaction of these two variables, providing support for lack of group differences to proceed with additional hypothesis testing. There were no Group differences on pre-FB RT, $F(3, 126) = .96$, $p = .41$. There was also no Problem main or interaction with Group. Last, the Groups did not differ on pre-FB FOR, $F(3, 126) = 1.24$, $p = .30$. However, there was a Problem main effect, $F(1, 126) = 13.17$, $p < .001$, $\eta^2_p = .10$. Conflict problems engendered lower confidence ratings (M = 5.09, SD = 1.21) than nonconflict confidence ratings (M = 5.26, SD = 1.19) for all groups. There was no Group by Problem interaction. These pretests suggest that the groups were more similar on this task than the base-rate task, so I will move on to the main analyses.

Main analyses. The remaining analyses were conducted to test the main study hypotheses for the conditional reasoning task. To ensure comparability with the base-rate neglect task, the same main 4 (Group) x 2 (Problem) x 2 (Pre/Post-Feedback) ANCOVAs were conducted with accuracy, RT, and FOR as separate dependent measures. Pre-FB scores were entered into these statistical models as covariates to control for performance on these

measures before the main feedback manipulation. Table 11 shows the means and standard deviations for the various effects described below.

Accuracy. The first test was conducted on accuracy of decisions on the task. With the adjustment of the covariate, there were no overall Group differences, $F(3, 125) = .11$, $p = .95$. Adjusted means were essentially slightly above chance. There was a marginal Problem effect, but the conflict problems still retained higher accuracy scores than nonconflict problems after the adjustment, $F(1, 125) = 3.59$, $p = .06$. There was a Feedback ME, $F(1, 125) = 65.46$, $p < .001$, $\eta^2_p = .34$, but this effect was practically uninterpretable because the adjusted means were essentially the same values. The covariate of pre-FB accuracy was a significant predictor of the model, $F(1, 125) = 233.95$, $p < .001$, $\eta^2_p = .65$. It interacted with Problem and Feedback. It appeared that participants in all groups did not improve as they progressed through the task and especially after the feedback information was given. Interestingly, conflict problems maintained higher accuracy in all groups throughout, especially when accuracy was measured across problems as opposed to overall task accuracy.

Response time. A similar story of effects on RT from the base-rate task is found here. Since there is an extensive practice effect on the first few problems carrying over to the remainder of the problem set, all groups get faster post-FB ($F(1, 125) = 21.54$, $p < .001$). With respect to the hypothesis that False-Low feedback would slow participants in this group relative to the other, it was not supported (the RTs are not the smallest rate reduction). Additionally, the False-High group did not increase in decision speed as a result of the feedback at the fastest rate (on both nonconflict and conflict problems, the Control group followed this trend, but the difference in RTs is not significant).

FOR. On the conditional reasoning task, the Feedback manipulation appeared to

modulate FOR as was intended by the manipulation. First, there is a Group main effect, $F(3, 125) = 2.92$, $p = .04$, $\eta^2_p = .07$. The False-High group had the greater FOR ratings, while the False-Low group had the lowest FOR ratings. In addition, there is a Group by Feedback interaction, $F(3, 125) = 3.02$, $p = .03$, $\eta^2_p = .07$. Further analysis into these two effects shows that, as hypothesized, the False-High feedback group gained confidence across problems, and both the False-Low feedback and True feedback groups rated lower confidence post-FB. However, since these differences in confidence are very small, the trends as hypothesized are not statistically reliable. As a manipulation check in this analysis, the control group, which did not receive any feedback, did not change confidence ratings after their rest period halfway through the task.

Conflict detection hypothesis. The prediction that conflict detection is efficient, routine, and universal (LIM) or a function of performance (TSM) was tested on the conditional reasoning task. Neither model was supported across Group, $r(127) = .15$, $p = .09$. The intercept for all participants was not different from zero and a negative value (-413 ms; $t(128) = -0.88$, $p = .38$), revealing that the majority of participants hovered near no RT difference of nonconflict RT and incorrect conflict RT. As shown in Figure 10, separating the groups revealed that only the False-Low feedback group had a significant positive relationship $r(29) = .41$, $p < .05$. The other three groups had unreliable correlations. In addition, the True feedback group's trend was negative and the best performer in this group only achieved an accuracy score of approximately .50.

Inhibition hypothesis. Overall, there was a negative relationship between conflict problem performance and RT differences between correct and incorrect judgments, but this correlation is not significant, $r(127) = -.05$, $p = .55$. The intercept to reflect the poorest

performance on the task was predicted to fall at less than one second—not enough to be different from zero in this regression model (977 ms, $t(128) = 1.23$. $p = .22$). Thus, this result reflects no clear case for inhibition across all participants. Separating the relationship by Group and the situation is extremely similar: none of the groups independently show inhibition ($-.25 < r\text{s} < .15$, $p\text{s} > .05$, see Figure 10).

Cognitive decoupling hypothesis. As with Experiment 2, there was no relationship between conflict problem accuracy and the RT difference between correct conflict responses and nonconflict responses, $r(127) = .04$, $p = .67$. The intercept rested at 564 ms, again under the threshold for additional processing revealed in this series of studies and other work (e.g., Mevel et al., 2014; Pennycook et al., 2015). Separating participants by group to tests this relationship revealed that the False-Low feedback group had a significant positive relationship ($r(29) = .42$, $p < .05$, see Figure 10, Panel C), and the True feedback group had a significant negative relationship as predicted by the TSM ($r(34) = -.38$, $p < .05$, see Figure 10, Panel D).

Feedback interaction with detection, inhibition, and cognitive decoupling. Table 12 shows the correlations of the three above indices with conflict problem performance before and after feedback. The majority of the correlations are not significant (flat), but the presence of feedback seems to change behavior patterns for participants. Notably, the relationship between the conflict detection index and conflict problem accuracy switched direction in the True feedback condition. In this group, Pre-FB, participants who made fewer errors showed a conflict detection effect (positive coefficient, TSM prediction supported); post-FB, however, participants in this group who in the end made fewer errors were faster and did not show the TSM-predicted positive slope. This makes sense: the feedback, however small (just a

percentage correct) allowed for faster responses after the rest period. Another explanation for this could be the practice effect that persists in all groups from the start of the task, which skews pre-FB RTs.

Conflict detection and metacognitive FOR. There was no relationship between metacognitive FOR and the cognitive detection index, $r(127) = .06$, $p = .50$, contrary to the hypothesis that these would be conversely related measures. Separating by groups revealed similar lack of relationships between conflict detection and FOR.

Results: Individual Differences and Time Series

Individual differences. Correlations between the performance variables discussed extensively and CRT, SAT, and NFC measures were conducted. As with Experiment 2, CRT scores correlated positively with SAT scores, $r(90) = .22$, $p = .04$. However, CRT scores did not correlate with NFC scores. Performance on the CRT was correlated with conditional nonconflict reasoning accuracy, $r(128) = .23$, $p = .007$. There were no other correlations present in the combined sample. Correlations were also computed by separating out Feedback Group. In the Control group, there were no notable or reliable correlations. In the False-High feedback group, CRT performance was positively correlated with conflict problem accuracy on the base-rate neglect task, $r(32) = .42$, $p = .04$. In the False-Low feedback group and the True feedback group, there were no reliable relationships between the measures and performance on either task.

The median split analyses from the previous two experiments were not conducted on this dataset due to the complexity of the design and the sample sizes for each Feedback group. It is reasonable to argue that the indices described above can differentiate bias-susceptible individuals from bias-resistant individuals. Bias-susceptible individuals appear to

err as a result of monitoring failures, while bias-resistant individuals appear detect conflict and choose the correct answer (decoupling).

Time series analysis. The same time series regressions from Experiments 1 and 2 were conducted for the two tasks in Experiment 3, with special attention paid to the feedback groups. Each regression analysis was conducted separately for the first half (before feedback) and the second half (after feedback) of each group.

Base-rate neglect. There was a large visible difference between conflict and nonconflict problems over time. This is the case for all the groups over the 50 problems. However, there were no noticeable differences between the trends by group. Prior to feedback, all groups had a declining trend on conflict problems. After feedback was presented, accuracy trends increased in all groups, with the sharpest (non-significant) change in the False-High group. However, this increase was larger than the trend for the False-Low feedback group. All groups had non-significant declining trends in FOR pre- and post-FB. The regression line for the False-Low group was constantly the lowest among all the groups. Again, the range was severely restricted on FOR ratings. RTs were not analyzed due to a consistent (across experiments) practice effect. There is nothing useful to glean from these trends.

Conditional reasoning. The conditional reasoning task, as I have described throughout Experiment 2 and this experiment, did not have the same robust dichotomy between conflict and nonconflict problems, because the problem types increase in their overall variability across time. It was difficult to decipher the trends on accuracy before and after feedback. The regression analyses did not reveal any effects. The FOR trend did show the mounting problem of the measurement of confidence through a 7-point Likert scale,

however. Large-scale effects were not present with very little change over time, which was more evidence that people tended to take each problem individually. Again, no RT time series were conducted using this method as all participants tended to speed up and large-scale trends would be influenced heavily by this.

Strategy change question. After the participants completed both tasks, they were asked if they had changed their strategy based on the feedback given. Participants also stated the strategies that they used. The Control group was only asked which strategy was used. Overwhelmingly, strategies included using stereotypes and probability on the base-rate task, whereas on the conditional reasoning task, participants suggested they relied upon logic or intuition (a synonym for believability, perhaps).

A one-way ANOVA was conducted to test Group on whether the participant changed their strategy after feedback was presented.[11] While the average for the False-High feedback group was the smallest (M = .30, SD = .47), there were no differences between the amount of participants who adopted a new strategy, $F(2, 96) = 1.29$, p = .28. Fewer participants in the False-Low feedback switched strategies (M = .45, SD = .51) than those in the True feedback (M = .49, SD = .51). Post-hoc comparisons corroborated the overall null effect. While a smaller proportion of participants in the False-High feedback group switched their strategy after feedback (less than one-third), not enough participants in the other two feedback groups switched strategies to show a distinct effect of the feedback manipulation. However, it is a promising trend.

[11] This test included 99 participants because the control group was not included. They did not receive feedback and were not asked about whether they changed their strategy.

Discussion

Experiment 3 was designed to examine the impact of feedback on the failure hypotheses discussed in DPTs (storage, monitoring, and inhibition failures, De Neys & Bonnefon, 2013). The central question was whether feedback interacted with these bias descriptions. For example, if answer fluency and FOR impacted subsequent T2 engagement, then utilizing a feedback manipulation designed to specifically address a person's sense of confidence in performance should work to modulate the instantiation of a monitoring failure. A monitoring failure as described by Thompson and colleagues (Thompson, 2009; Thompson et al., 2011) is the failure of a metacognitive feeling. A monitoring failure as described by De Neys and Bonnefon (2013) is the failure to recognize that the situation requires formalized knowledge; Pennycook et al.'s (2015) TSM takes this further by defining a monitoring failure as simply a failure of conflict detection. If one assumes that the FOR and behavioral conflict detection (by RT) are related or converging measures of the same cognitive attribute, then a manipulation through feedback should hope to extinguish this bias. Three groups in this experiment received some form of feedback, while a control group acted as a baseline. One group received feedback intended to bolster confidence (False-High), another group received feedback intended to reduce confidence (False-Low), and final feedback group received feedback intended to reflect actual performance (biased decisions prior to feedback would dictate the effect of feedback; True feedback group). These groups were inserted into Experiment 2's procedure where participants completed the base-rate neglect and the conditional reasoning tasks.

Base-rate neglect. On the base-rate neglect task, overall accuracy effects replicated, but RT and FOR effects were difficult to parse. Overall, the feedback group manipulation did

very little interacting with the robust existing effects (replicated in the Control group) on the base-rate tasks. There were promising trends in the False-Low group with respect to FOR: this group has the lowest rated FOR after feedback was introduced. The inferred T2 engagement was not reflected in RT slowing or increased accuracy versus the other 3 groups, however.

With respect to the model predictions on behavioral individual differences, there was added support for the TSM (Pennycook et al., 2015) overall and within each group. Thus, participants who made the fewest errors in each group spent more time detecting conflict than people who made the most errors. This is similarly the case in Experiments 1 and 2. The inhibition prediction of the LIM was not supported (De Neys, 2012, 2014a, 2014b). Last, the TSM postulation of decoupling for the most-biased participants was supported: they spent more time working out correct responses than participants who made the fewest errors. These three indices did not interact with the feedback manipulation, however (there were no significant relationship changes for any group).

Conditional reasoning. On the conditional reasoning task, two interesting results were observed. Though there was an accuracy effect of feedback, I would hazard to say that there is no practical difference across groups before and after feedback was given, as the adjusted means were only .007 units apart. The likely issue here was that the combination of nonconflict and conflict problems were less clear than in Experiment 2. The second interesting effect was that there was a Group by Feedback interaction on FOR ratings. These ratings decreased in the False-Low and True feedback groups after feedback, but increased in the False-High feedback group. As with the base-rate task, this effect offered a glimmer of a trend, but accuracy and RT measures were not differentially impacted as predicted.

The behavioral indices on this task were as muddied here as in Experiment 2; overall there was no support for either the TSM or LIM. There were no clear patterns to interpret averaging across feedback groups or for the individual feedback groups.

Additional considerations. First, response times were difficult to analyze directly within the main ANCOVAs on both tasks. Time series analyses show a large decrease in RTs across trials. Pre-FB RT averages were always slower than post-FB averages. The lack of training problems prior to the main task imparted a significant practice effect that persisted in all groups and across each of the experiments. In this case, coupled with the restricted range of the recorded FOR ratings, inferring T2 engagement in this overall study is premature.

Second, Experiment 3 added to the evidence that the LIM (De Neys, 2012, 2014a, 2014b) is not sophisticated enough to account for individual differences in response times and behavior. It is clear that monitoring failures defined as errors of conflict detection are more predictive of a person's ability to do reasoning/judgment tasks than inhibition failures. Though the data were inconclusive, the TSM (Pennycook et al., 2015) model was primarily supported. Additionally, Experiment 3 added to the conclusions of Experiment 2 that both models cannot account for the tasks that extend beyond the social inference heuristics of Kahneman and Tversky (1973), such as conditional reasoning task. Conditional reasoning requires some formalized knowledge in addition to some logical intuitions (modus ponens) that social inferences rarely need (if the heuristic is strong enough). The base-rate neglect task used here is qualitatively distinct from the conditional reasoning task used here, and the data in this series of experiments support this distinction. The models (LIM, TSM) are task-specific to base-rate neglect.

Chapter V.

General Discussion

The central question of the present dissertation was source of errors people make on reasoning and judgment decisions and whether people know that they are biased. It addressed this question by investigating three major pieces within a Dual Process Theories framework, namely (1) the role the conflict detection and monitoring mechanism, as described by De Neys (2012, 2014a) in the LIM, and as described by Pennycook et al. (2015) in the TSM; (2) the scope of these models and if predictions can be applied to different tasks than those from which they were developed; (3) and whether the cognitive failures could be mitigated by offering a feedback intervention.

Theoretical Implications: Conflict Detection and Resolution

Conflict detection and eventual resolution (whether achieving the "correct" or "incorrect" answer to a given judgment or reasoning problem) is an essential function of any dual process framework (De Neys, 2012, 2014a, De Neys & Glumicic, 2008; Evans, 2007, 2009; Pennycook et al., 2012; Pennycook et al., 2015) because it identifies *how* and *when* a person would utilize T1 responses over T2 responses, or vice versa. Early processing models (serial and parallel) failed to account for how conflict was handled by the two processing types (e.g., Epstein, 1994; Kahneman & Frederick, 2002; Sloman, 1996), which has engendered criticism to the overall dual process framework (e.g., Keren, 2013; Kruglanski & Gigerenzer, 2011). Moreover, a specific description of the conflict mechanism is required to support the case of dual processing vs. unimodal or continuum processing (Osman, 2004). Two recent models of conflict detection within DPT have garnered considerable attention (De Neys, 2014a). The LIM postulates that conflict detection is a universal trait, routinely

utilized, and efficient/quick (De Neys & Glumicic, 2008; De Neys, 2012). The TSM postulated that detection was not universal, an individual difference, but for those who utilize it, can be quite quick (Pennycook et al., 2015). Using the description of biases by De Neys and Bonnefon (2013), the LIM claimed that errors were the result of inhibition failures; the TSM claimed errors were due to monitoring failures.

Monitoring or inhibition failures? The results from three related experiments are possibly consistent with the TSM, but there are a number of qualifications that must be made. Overall, conflict detection was faulty for those participants who tended to make many biased responses. The group averages promoted an overall conflict detection effect, but once individuals were compared against this group effect, it was clear that monitoring failures, and not inhibition failures, were the cause of this bias. The LIM's postulation that conflict detection should be a flat relationship with a positive intercept was not supported. In terms of T1 and T2 processing, the line drawn between the two is clear and the TSM is better suited to account for the present data. According to the TSM, in Stage 1 of the thinking processing, a reasoner utilizes T1 processing—the initial responses are generated based on intuition (Evans, 2007), feelings of rightness (discussed in more detail below; Thompson, 2009), answer fluency (Thompson et al., 2011), or in response to cognitive load (De Neys, 2006) or time pressure (Evans & Curtis-Holmes, 2005), and generated rapidly (in milliseconds). If there is no conflict between various initial responses, such as when a base-rate problem cues the same answer from both the actual numbers and the stereotypic information, decisions generally proceed within milliseconds from the generation of responses. This is a good baseline for all responses, but particularly for assessing the engagement of T2. Now, in the case of conflict in the initial responses, a resolution must be made. In this simplified

dichotomy of responses, the heuristic response is at odds with the probabilistic/normative response. The conflict monitoring mechanism reflects additional processing. This additional processing, whether a person decides to choose the heuristic response or the normative response, is a reflection of T2 engagement (Pennycook et al., 2015). What T2 does in the moments after the conflict is initially detected is identified by models as inhibition/inhibition failure (LIM) or cognitive decoupling/rationalization (TSM).

Though the TSM does not specifically address inhibition failure (it was classified as part of a cognitive decoupling process), the results across studies are inconsistent with an LIM perspective that describes the function of T2 as an inhibitor. The negative correlation between inhibition RTs and conflict problem accuracy indicated that the fewer errors made on these problems generally led to quicker responses. Thus, the better performers on the two thinking tasks generally spent less time making correct responses than making incorrect responses. It makes sense that the participants who made the most errors had to inhibit the incorrect response on a limited number of trials, but the LIM would predict that the inhibition processing times ultimately increase as a person makes more correct decisions, as T2 becomes an active inhibitor, which would result in a positive slope for this relationship. The LIM is not sophisticated enough to readily explain the data; however, conceptual pieces of the model are not lost in the later formulation of the TSM.

The final piece to the biases question is the predictive power of the indices to classify individuals based on their conflict problem performance. First, performing a median split on conflict problem accuracy averages created two groups. Below-median individuals were classified in the bias-susceptible group and above-median individuals were classified in the bias-resistant group. Using conflict accuracy was by far the best classifier; however, this is

unsurprising. This does not solve the issue as the main source of errors on these reasoning and judgment tasks. Using the conflict detection index and the inhibition index as the sole predictors, a discriminant function analysis was conducted on all datasets and tasks. Across all experiments, the conflict detection index was the stronger predictor between the two, especially on the base-rate neglect task. Based on just these two predictors, and primarily the detection index, participants were correctly classified in the bias-susceptible and bias-resistant groups greater than 53% of the time in Experiments 1 and 2 (this lower bound was the conditional reasoning task of Experiment 2). RT differences between nonconflict baseline and incorrect conflict responses (conflict detection index) identified group placement better than the RT difference between correct and incorrect conflict responses (inhibition index). This represents further evidence to support monitoring failures as the major source of bias across tasks (De Neys, 2014a; Pennycook et al., 2012; Pennycook et al., 2015).

Though monitoring failures are the observed source of errors across the base-rate experiments, and somewhat on the conditional reasoning task, inhibition failures cannot be ruled out as a potential source for some people. De Neys (2012, 2014a) does not label what an inhibition failure may involve, judging it as a merely a failure of the structure of the RLPFC (De Neys et al., 2008) to stop a compelling, strong initial response (Botvinick et al., 2004). Pennycook and colleagues lump inhibition failures into decoupling failures (Pennycook et al., 2015). Whether these failures are a product of rationalization is still unknown, though verbal protocols might be able to detect this specific distinction (such as the methodology of De Neys & Glumicic, 2008). In a two-alternative forced-choice paradigm, a person is either correct or incorrect, which leaves little room for interpretation of a mechanical failure (inhibition) vs. a psychological failure (rationalization). The results of

these studies did not support the cognitive decoupling hypothesis, *per se*, but showed that bias-resistant individuals sometimes make mistakes. In fact, it seemed from these results that a positive decoupling relationship makes more sense than a negative relationship, because if a person is spending more processing time to detect conflict, then they should take additional time to decouple, thus increasing time above a baseline (which is what is predicted by the inhibition index account of the LIM). Whether this leads to a correct or incorrect answer depends entirely on the definition of decoupling, which broadly defined, would include rationalization (Evans & Stanovich, 2013). Taken together, inhibition failures may just be rationalizations to choose the compelling initial response. This explains those bias-resistant individuals on both tasks making the occasional errors.

Storage failures? While it is entirely plausible that some people have storage failures (Stanovich, 2009)—they do not have the requisite knowledge to perform validity evaluations—it is difficult to determine that this source of error is behaviorally distinct from monitoring failures. Both types of failures are marked by low accuracy and fast response times on nonconflict and conflict problems. The dimension that might separate people into one failure source or the other on a given task would be confidence/FOR, or the factor that would show that they do not know that they are making decisions in a biased way. To know if this is true, a storage failure may be marked by low confidence (an answer of this nature should be "Guessing" [1] on the scale used in these experiments). According to the metacognitive monitoring theory (Thompson, 2009; Thompson et al., 2011), this low confidence should signal T2 engagement. However, in this case, T2 would not engage on these problems because there is nothing to analyze, because there is no formal knowledge to retrieve and evaluate. Conversely, a monitoring failure might be marked with a high FOR—

no detection that the problem or situation requires T2 engagement, but confidence of the initial response. Conceptually, it makes sense to divide behavioral responses in this way. However, the data presented here do not support this distinction. Storage failures were not observed here: incorrect response FOR was not correlated with RT on those responses for conflict base-rate problems (r = .01) and this was similar for incorrect nonconflict problem RT and FOR (r = .05). This is not surprising, as the base-rate neglect task does not necessarily lend itself to a storage failure explanation. However, on the conditional reasoning task, where the explanation is plausible for errors, a similar case is observed: incorrect RTs on both nonconflict and conflict problems were not correlated with the respective FOR values (they were actually negatively correlated, rs = -.08). Thus, I am confident that the monitoring of the conflict detection mechanism is the weak link for the majority of individuals in in the cognitive framework of dual processing routes.

FOR. Metacognitive FOR (Thompson, 2009; Thompson et al., 2011) offered an additional perspective and measure to determine the role of T2 engagement in both the base-rate neglect task and conditional reasoning task. It was a self-report measure of confidence that participants would record following each question they encountered. In Experiment 2, the role of FOR was mixed. At the group level, problem type and response interacted to produce FOR effects on both tasks. This suggested that the nature of the problems, whether they were nonconflict or conflict, changed the way participants evaluated their confidence in their answers. This sounds like strong support for the measure and the conceptual link that low FOR signals the need for T2 engagement or that high FOR does not signal this engagement. However, the remainder of the findings muddies these conclusions. As a covariate, it did not affect RTs on either task. This is likely due to the temporal ordering of

measurement whereby the response followed by the confidence rating was not sophisticated enough to capture the milliseconds worth of processing. This ordering could have also led to a minor hindsight bias or response fence-sitting. Further evaluation of the confidence scale will be addressed in a later section of limitations. As a converging measure of T2 engagement with conflict monitoring postulated by the TSM, there were weak correlations. Conceptually, FOR should have a negative relationship with conflict monitoring: high FOR should reduce the need for conflict monitoring, and errors should be relatively fast with a high confidence rating. Conversely, low FOR should signal greater monitoring within the individual, increasing RT from baseline and possibly leading to a correct answer (see Thompson & Johnson, 2014).

In Experiment 3, a manipulation of FOR was attempted. If the presence of feedback interacted with FOR, which in turn would interact with accuracy and RTs, then large group differences should be observed. This was not the case in this experiment. There were trends that appeared, which suggest some methodological considerations, such as measurement of FOR or problem presentation changes for increased power. Though these are minor findings to support the overall impact of FOR in DPT, they should not be disregarded completely. It is clear more work is needed with stronger methodology. Future research should also address if FOR can be untangled from the neurophysiological evidence on conflict detection, perhaps utilizing the rethinking methodology of Thompson et al. (2011).

Task-specific or task-general models? Both the LIM and TSM were developed using the base-rate neglect task (Kahneman & Tversky, 1973; Tversky & Kahneman, 1974), which has shown robust accuracy (choosing the base-rate answer) effects for decades (e.g., De Neys & Glumicic, 2008; Pennycook et al., 2012). Robust RT effects have been shown for

almost a decade when utilizing extreme base rates (e.g., 997 vs. 3; De Neys & Glumicic, 2008). It was important for these two models to be tested outside of base-rate neglect, on something other than a problem that invokes a social inference (even conjunction fallacy problems fall under this category; De Neys, 2014b). In recent DPT investigations, conditional reasoning has been used (Thompson et al., 2011), as well as denominator neglect/ratio bias (Mevel et al., 2015; Thompson & Johnson, 2014) and categorical syllogistic reasoning (De Neys & Franssens, 2009; Thompson & Johnson, 2014). Conditional reasoning was a reasonable choice from these tasks for replication and extension purposes. Conditional reasoning, especially inferences about definitional or causal rules, does not involve making inferences about people, merely the connection between objects or outcomes, and is therefore qualitatively distinct from base-rate neglect. Furthermore, though general explanations of DPTs were developed in syllogistic reasoning (e.g., Evans et al., 1983) or other reasoning tasks, the two models were created with base-rate neglect, and so their specific predictions have not been evaluated by other researchers or tasks in the extant literature since the models' development.

Experiments 2 and 3 contained both the base-rate neglect and conditional reasoning tasks for a direct comparison with the same participants. Perhaps the only finding that carries through both tasks is that monitoring failures are the likely cause of bias, but even this effect is unclear across tasks, i.e., the conditional reasoning task. RT differences were extremely close on the conditional reasoning task, which do not match the larger effects observed on the base-rate neglect task. The data support the conclusion that the LIM cannot describe the observed patterns of behavior. Additionally, I am reticent to suggest that the TSM applies across tasks because the effects are extremely weak on the conditional reasoning task. It may

actually be the case that bias on this task reflects storage failures (Stanovich, 2009) than monitoring failures, which is something that the TSM cannot parse in behavioral responses.

Experiment 3 further exacerbated the distinction between the two tasks. The feedback manipulation was stronger on the conditional reasoning task, which ultimately showed the glimmer of support for manipulating confidence in dual process methodology. Feedback minimally impacted the base-rate neglect task. Overall, responses on conditional reasoning problems reflected different processing that could not be accounted for by the two models tested in this dissertation that were readily available for base-rate responses.

I do not suggest that these DPTs are incompatible with conditional reasoning. The conclusion I draw from these data and this series of studies is that the conflict detection and monitoring models developed within base-rate neglect do not transfer to a qualitatively different task. I suggest that these models are task-specific. Further tuning is necessary to encapsulate the spectrum of reasoning tasks into a dual process framework.

Thinking dispositions, cognitive ability, and conflict detection. What do the results suggest for the inclusion of thinking dispositions or cognitive ability in identifying individual differences? This dissertation has some promising findings to add to the discussion of cognitive engagement with these tasks (Evans, 2007); Experiment 1 had typical correlations between performance on the base-rate neglect task and the CRT (Frederick, 2005), the AOT (Stanovich & West, 1997), and NFC (Cacioppo et al., 1984; see Swan & Revlin, 2015, for a review). However, with the exclusion of the AOT in Experiments 2 and 3, very little could be gleaned from thinking dispositions and the relationship with task performance and cognitive traits. The strong correlations from the first experiment disappeared in these two samples. There were some correlations between the RT indices and disposition and ability that were

interesting: positive relationships for conflict detection and NFC and CRT scores in Experiment 1 (rs > .31, ps < .01), suggesting that as a participant was more interested in complex cognitive activities (NFC) or more reflective in the face of intuitive answers (CRT), more time was spent on detection of conflict across problems. In addition, there was a small positive correlation of the conflict detection index with cognitive ability (SAT scores), suggesting that people with a stronger ability to perform well actually did do well and made fewer monitoring errors (e.g., Stanovich & West, 2000; Toplak, West, & Stanovich, 2014).

Some of these effects disappear in Experiment 2. While NFC scores are positively related to conflict detection on the conditional reasoning task, they are not on the base-rate task, contrary to Experiment 1. SAT scores predict conflict detection on the base-rate task but not on the conditional reasoning task—further evidence that these two tasks are distinct. Similar effects appear in Experiment 3.

These mixed findings show the lack of clarity with these reasoning and judgments task as and the role of cognitive ability and thinking disposition. It might be the case that SAT scores are not a good indicator of ability, as previous literature maintained (Stanovich & West, 2000), and that thinking dispositions are too variable (Svedholm-Häkkinen, 2014) and need further validation (self-report measures may generally have this issue). Further research is needed to full incorporate and address these individual difference measures into the failure individual differences discussed here.

Three-Stages Model Modifications

In an effort to refine conflict detection models and integrate them with the data and results explored here, I offer some important changes to the TSM model (Pennycook et al., 2015) for further exploration in future research. As Figure 11 illustrates, T1 and T2

processing remain serialized. The observable response outputs from the two processing routes remain quantifiably distinct. The line at the bottom of the Figure represents time and its progression through the three stages. The three stages are marked within the T1 and T2 processing routes. First, once a problem or situation is presented, initial responses are generated based on FOR, fluency, or other indicators (Thompson et al., 2011). In the tasks used frequently by various researchers (e.g., De Neys & Glumicic, 2008; Pennycook et al., 2012; Pennycook et al., 2015; Thompson & Johnson, 2014; Thompson et al., 2011) all problems have the potential to generate multiple responses. I have used the LIM's structure by limiting this to two intuitive responses (De Neys, 2012): an initial response of a heuristic intuition (IR_H) or an initial response of a logical or probabilistic intuition (IR_L).

The major change I make to the model is to explicitly include the conflict monitoring stage (Stage 2) in T2 and not within T1 (this delineation is not clear in the original TSM, see Figure 2). There is an important distinction in this stage regarding the timing of conflict detection. I argue that based on the data collected in three separate experiments, conflict detection is a gradual sub-process of T2 that takes longer than milliseconds, in contrast with how the timing is described in TSM. However, if there is no conflict present or detected, initial saliency-based responses still do occur within milliseconds after completing the initial reading of the problem. This accounts for nonconflict responses and undetected, incorrect conflict responses (IR_H), especially on both tasks but maintains these decisions as functions of T2. This is because working memory will contain the decision made and a person can report exactly what decision they have made (Evans, 2009; Evans & Stanovich, 2013). In addition, the response output from an undetected conflict in a two-alternatives problem could be a correct answer (IR_L). As particularly evidenced by a number of participants (44% of the

sample) on the conditional reasoning task in Experiment 2, many correct responses were actually made quicker than the nonconflict baseline (decoupling index), which could be accounted for by labeling these responses to either a saliency-based strategy, such as the structure of the argument was more apparent than the necessity/sufficiency of the conditional rule, or just basic guessing. The original TSM fails to incorporate responses that reflect this possibility (Pennycook et al., 2015).

On problems where the two responses conflict, once conflict is detected between IR_H and IR_L, T2 must now continue to a decision between the two alternatives. This initiates Stage 3, under my umbrella term of cognitive decoupling. This defining feature of cognitive decoupling (Evans & Stanovich, 2013) leads to a correct or incorrect answer through working memory deliberation. A correct answer is interpreted as a normalization (use of normative logic) or inhibition of the IR_H, which leads to the selection of IR_L. An incorrect answer reflects rationalization (as described by Pennycook et al., 2015), whereby both answers were weighed by T2, but ultimately confidence or other factors indicate a preference for IR_H. The relative distance between the rationalization response and the normalization response support the idea that incorrect answers on conflict problems do tend to take longer than correct, normalized answers. The bias-resistant participants who tended to spend much deliberation on incorrect answers on both tasks support this latter explanation.

With the TSM model expressed in this modified way, one can clearly identify the bias divergences described De Neys and Bonnefon (2013). First, if a person suffers from storage failures, the path to a quick decision is generally through the IR_H path in the model. If conflict monitoring is the source of errors, bias-susceptible individuals are separated very early from bias-resistant individuals (across problems), and guessing can be determined by

examining the detection index vs. the decoupling index for correct answers (such as if the decoupling index value is below the nonconflict baseline and accuracy is low, this is likely a guessing situation) Further divergence can occur after conflict is detected within bias-resistant individuals: normalizers/inhibitors (participants who tend to get problems mostly correct) or rationalizers (who get more incorrect than normalizers and might be more bias-susceptible). The important thing to remember with the original TSM and this modified model is that it exists on a problem-to-problem basis, where a person of any description could come to a correct or incorrect decision at any time. This ultimately becomes the limiting factor of the model: there are only two alternatives.

The purpose of these modifications was not to discount the TSM completely, but to allow it stronger explanatory power for the present data. However, these modifications need further examination in the base-rate neglect task and the conditional reasoning task, as well as other social inference tasks (Kahneman & Tversky, 1973) and traditional reasoning tasks, such as linear reasoning or categorical syllogistic reasoning. It is possible that further tuning is required. Incorporation of these modifications into similar methodologies is also warranted investigation, especially with the methodological limitations described in the next section.

Limitations

Many of the above conclusions and evaluation of the theoretical implications need to be qualified within some limitations that may have influenced the observed results. First, while there were some promising FOR effects, it is appropriate to mention the methodological issues with the scale used to gather FOR ratings. I do not challenge the effects observed on this scale; they replicate the differences that Thompson et al. (2011) and Thompson and Johnson (2014) observed and were later used in theoretical justifications for

the measure. However, this does not absolve it from critical analysis. For example, a scalar difference of .20 between nonconflict and conflict problems (Thompson et al., 2011, p. 123) is hardly descriptive in practical terms. What do these intra-scale values actually represent (even when they are averaged across multiple trials)? It might have to do with the interpretation of the anchors on the scale. Included with the numbers are three verbal anchors, with two for the endpoints and one for the midpoint. The high endpoint reads "Certain I'm Right", while the low endpoint reads "I'm Guessing" (Thompson et al., 2011). The remainder of the scale reflects gradations between guessing and certainty. However, the midpoint represents the problem; it reads "Fairly Certain". The scale was slightly modified in Thompson and Johnson (2014) whereby "I'm Guessing" was changed to "Doesn't Feel Right at All" to better reflect the "feeling" part of FOR. However, this change did little to dynamically change responses on the scale and small mean differences were observed across tasks (Thompson & Johnson, 2014). Moreover, the time series analyses in Experiments 2 and 3 show that the majority of participants, on average, were generally more confident in their answers than "fairly certain". However, it is reasonable to suggest that the interpretation of the midpoint is open and malleable. In principle, that does not disqualify the scale or the verbal anchors, but it does call into question the responses when behavioral measures indicate guessing. Additionally, the resultant restricted range has consequences for FOR's relationship with the other variables in question. Conclusions drawn from this scale should be taken and made with caution. Future research of metacognitive FOR would benefit from a methodological shift from this scale of confidence to a more reliable and validated measure of confidence (or utilize a collection of measures).

A related limitation to the FOR/confidence scale was its placement within the general

procedure. While participants were answering the questions within each task, their metacognitive sense of confidence was always *post hoc* to the behavior of generating the answer. Some people may have had difficulty introspecting in this way. This necessary ordering issue may have impacted ratings of confidence, but reviewing response times for confidence ratings show that the majority of responses on the scale for each problem within a task were usually less than a second—the initial response to the question was weighed heavily, reflecting the theorized implicit mechanisms of FOR, but the selection of the confidence rating did not generally receive much deliberation or introspection. If the latter is the case, then the limitations discussed in the previous paragraph are far more important for interpretation of the concept of FOR.

The other methodological limitation to this dissertation reflects an oversight in Experiment 3's design. In the general design, random assignment was used to prevent confounds in the between-subjects group variable, and complete randomization was used for the individual problems within each task to prevent order effects within the tasks (problem ordering), but the two tasks were not counterbalanced to control for large-scale carryover effects from the first task to the second. This control limitation is especially important in the two false feedback conditions (False-High and False-Low groups), where the participants were told essentially the same information for both tasks. For each participant, the base-rate neglect task always came first, rather than half receiving the conditional reasoning task first. Feedback on the base-rate task may have affected performance in the pre-FB portion of the conditional reasoning task. This may have initiated a feeling of failure or overconfidence before the second task began. This is a possible confound within the results of the conditional reasoning task. However, if the effects were truly damaging to the overall interpretation of

the findings, there would likely be anomalies in the Feedback groups vs. the Control group, since the latter did not receive feedback and would therefore not reflect any problem with task carryover. This is not the case in the data, as the pre-FB analysis did not reveal group differences. Though the data may or may not support a carryover effect, it cannot be ruled out because I cannot compare participants who received different tasks first and test for the null effect (no differences).

Conclusion

The purpose of this dissertation was to investigate the general question related to whether people know they are biased, how errors manifest on classic reasoning and judgment tasks, and importantly, if DPT models are task-general explanation for cognitions and behaviors. First, despite the limitations outlined above, it is clear that conflict detection and resolution is an individual difference, as postulated by the TSM. Conflict monitoring is essential for detecting multiple cued responses by a given problem, context, or situation. It is also not equally efficient in all folks who have active monitoring mechanisms. The link between metacognitive Feelings of Rightness is fuzzy, but with some methodological changes, future investigation could be a promising endeavor.

Second, some models developed within a given task suffer from difficult boundaries. The two recent models described herein represent such a case: the complexity and qualitative differences of conditional reasoning at best reflect observationally different processing from base-rate neglect. The modifications to the TSM attempt to address these processing idiosyncrasies and future studies are planned to incorporate these new predictions.

While it is not the intention of this manuscript to speak directly to the general criticisms of DPTs, it is somewhat clear that these theories may be handy or nifty

explanations of cognitions, as many dichotomies share, but that they may ultimately be too simplistic to capture the range of cognitions and entirety of the vast cognitive architecture. This pause may not be at the level of some researchers' opinions (e.g., Gigerenzer, 2009; Kruglanski & Gigerenzer, 2011), but it should signal to DPT researchers to consider efforts for over-simplification.

It is, however, safe to assume that humans may be biased and poor reasoners because their cognitive architecture is designed for efficiency, and this so happens to prevent many from even recognizing that they are biased.

References

Aminoff, E. M., Clewett, D., Freeman, S., Frithsen, A., Tipper, C., Johnson, A., … Miller, M. B. (2012). Individual differences in shifting decision criterion: A recognition memory study. *Memory & Cognition*, *40*(7), 1016–30. http://doi.org/10.3758/s13421-012-0204-6

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*(12), 539–46. http://doi.org/10.1016/j.tics.2004.10.003

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307. http://doi.org/10.1207/s15327752jpa4803_13

Chen, S., & Chaiken, S. (1999). The Heursitic-Systematic Model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process Theories in Social Psychology* (pp. 73–96). New York: Guilford Press.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*, 317–370.

De Neys, W. (2006). Dual processing in reasoning. *Psychological Science*, *17*(5), 428–433.

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

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

De Neys, W. (2014b). Feedback and heuristic bias: I knew it all along. Paper presented at the

55th Annual Meeting of the Psychonomic Society. Long Beach, CA.

De Neys, W., & Bonnefon, J.-F. F. (2013). The "whys" and "whens" of individual differences in thinking biases. *Trends in Cognitive Sciences*, *17*(4), 172–178. http://doi.org/10.1016/j.tics.2013.02.001

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition*, *113*(1), 45–61. http://doi.org/10.1016/j.cognition.2009.07.009

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

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

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

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect we are biased. *Psychological Science*, *19*(5), 483–489.

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

Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. http://doi.org/10.1016/j.tics.2003.08.012

Evans, J. St. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321–339. http://doi.org/10.1080/13546780601008825

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

Evans, J. St. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. St. B. T. Evans & K. Frankish (Eds.), *In Two Minds: Dual Process and Beyond* (pp. 33–54). Oxford: Oxford University Press.

Evans, J. St. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning*, *18*(1), 5–31. http://doi.org/10.1080/13546783.2011.637674

Evans, J. St. B. T. (2013). Dual-process theories of reasoning: Facts and fallacies. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 115–133). New York: Oxford University Press.

Evans, J. St. B. T. (2014). Two minds rationality. *Thinking & Reasoning*, *20*(2), 129–146. http://doi.org/10.1080/13546783.2013.845605

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.

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. http://doi.org/10.1080/13546780542000005

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.

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. http://doi.org/10.1177/1745691612460685

Evers, C., de Ridder, D. T. D., & Adriaanse, M. A. (2009). Assessing yourself as an emotional eater: Mission impossible? *Health Psychology*, *28*(6), 717–25. http://doi.org/10.1037/a0016700

Frankish, K., & Evans, J. St. B. T. (2009). The duality of mind: An historical perspective. In J. St. B. T. Evans & K. Frankish (Eds.), *In Two Minds: Dual Process and Beyond* (pp. 1–29). Oxford: Oxford University Press.

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

Gigerenzer, G. (2009). Surrogates for theory. *APS Observer*, *22*(2), 21–23.

Goel, V., Buchel, C., Frith, C., & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage*, *12*(5), 504–514. http://doi.org/10.1006/nimg.2000.0636

Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition*, *87*, B11–B22. http://doi.org/10.1016/S0

Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 62, pp. 33–58). Waltham, MA: Academic Press. http://doi.org/10.1016/bs.plm.2014.09.002

Hutchison, J., & Martin, D. (2015). The evolution of stereotypes. In V. Zeigler-Hill, L. L. M. Welling, & T. K. Shackleford (Eds.), *Evolutionary Perspectives on Social Psychology* (pp. 291–301). New York: Springer.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and*

Biases: The Psychology of Intuitive Judgments (pp. 49–81). New York, NY: Cambridge University Press. http://doi.org/10.1017/CBO9780511808098

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

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. http://doi.org/10.1037/h0034747

Keren, G. (2013). A tale of two systems: A scientific advance or a theoretical stone soup? Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*(3), 257–263. http://doi.org/10.1177/1745691613483474

Kruglanski, A. W. (2013). Only one? The default interventionist perspective as a unimodel--Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*, 242–247. http://doi.org/10.1177/1745691613483477

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review*, *118*(1), 97–109. http://doi.org/10.1037/a0020762

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

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, *27*(2), 227–237.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. http://doi.org/10.1037/0033-

295X.84.3.231

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences*, *5*(8), 349–357. http://doi.org/10.1016/S1364-6613(00)01699-5

Oliver, J. (Writer) & Pennolino, P. (Director). (2016). Scientific studies [Television series episode]. In T. Carvell, J. Oliver, J. Taylor, & J. Thoday (Executive Producers), *Last Week Tonight with John Oliver*. New York: Home Box Office, Inc.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review*, *11*(6), 988–1010.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition*, *42*(1), 1–10. http://doi.org/10.3758/s13421-013-0340-7

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, *124*(1), 101–106. http://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*(October), 34–72. http://doi.org/10.1016/j.cogpsych.2015.05.001

Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review*, *19*(3), 528–534. http://doi.org/10.3758/s13423-012-0249-3

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), 544–554. http://doi.org/10.1037/a0034887

Petty, R. E., & Wegener, D. T. (1999). The Elaboration Likelihood Model: Current status and controversies. In S. Chaiken & Y. Trope (Eds.), *Dual-process Theories in Social Psychology* (pp. 41–72). New York: Guilford Press.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*(1), 133–48.

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

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. St. B. T. Evans & K. Frankish (Eds.), *In Two Minds: Dual Process and Beyond* (pp. 55–88). Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199230167.003.0003

Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. New York: Oxford University Press.

Stanovich, K. E., & Toplak, M. E. (2012). Defining features versus incidental correlates of Type 1 and Type 2 processing. *Mind & Society*, *11*(1), 3–13. http://doi.org/10.1007/s11299-011-0093-6

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*(2), 342–357.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–65; discussion 665–726.

Stollstorff, M., Vartanian, O., & Goel, V. (2012). Levels of conflict in reasoning modulate right lateral prefrontal cortex. *Brain Research*, *1428*, 24–32.

http://doi.org/10.1016/j.brainres.2011.05.045

Svedholm-Häkkinen, A. M. (2015). Highly reflective reasoners show no signs of belief inhibition. *Acta Psychologica*, *154*, 69–76. http://doi.org/10.1016/j.actpsy.2014.11.008

Svedholm, A. M., & Lindeman, M. (2013). The separate roles of the reflective mind and involuntary inhibitory control in gatekeeping paranormal beliefs and the underlying intuitive confusions. *British Journal of Psychology*, *104*, 303–319. http://doi.org/10.1111/j.2044-8295.2012.02118.x

Swan, A. B., & Revlin, R. (2015). Inhibition failure is mediated by a disposition toward flexible thinking. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of 37th Annual Meeting of the Cognitive Science Society* (pp. 2314–2319). Austin, TX: Cognitive Science Society.

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory & Cognition*, *22*(6), 742–758. http://doi.org/10.3758/BF03209259

Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. S. B. T. Evans & K. Frankish (Eds.), *In Two Minds: Dual Process and Beyond* (pp. 171–195). Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199230167.003.0008

Thompson, V. A. (2013). Why it matters: The implications of autonomous processes for dual process theories—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science*, *8*(3), 253–256. http://doi.org/10.1177/1745691613483476

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

Thompson, V. A., & Morsanyi, K. (2012). Analytic thinking: Do you feel like it? *Mind &*

Society, *11*(1), 93–105. http://doi.org/10.1007/s11299-012-0100-6

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107–40. http://doi.org/10.1016/j.cogpsych.2011.06.001

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

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*(2), 147–168. http://doi.org/10.1080/13546783.2013.844729

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

Tables

Table 1

List of Acronyms Appearing in Manuscript

Acronym	Full Term
DPT	Dual Process Theories
JDM	Judgment and Decision-Making
ACC	Anterior cingulate cortex
RLPFC	Right lateral prefrontal cortex
LIM	Logical Intuition Model
TSM	Three-stages Model
FOR	Feeling of Rightness
CRT	Cognitive Reflection Test
NFC	Need for Cognition
AOT	Active Open-minded Thinking
MP	Modus Ponens
MT	Modus Tollens
AC	Affirming the Consequent
DA	Denying the Antecedent

Table 2

Clusters of Attributes Associated with Dual-Process Theories for Each Processing Type

Clusters	Type 1	Type 2
1. Consciousness	Unconscious (preconscious)	Conscious
	Implicit	Explicit
	Automatic	Controlled
	Low effort	High effort
	Rapid	Slow
	Low capacity	High Capacity
	Default process	Inhibitory
	Holistic, perceptual	Analytic, reflective
2. Evolutionary	Old	Recent
	Evolutionary rationality	Individual rationality
	Shared with animals	Uniquely human
	Nonverbal	Linked to language
	Modular cognition	Fluid intelligence
3. Functional Characteristics	Associative	Rule-based
	Domain-specific	Domain-general
	Contextualized	Abstract
	Pragmatic	Logical
	Parallel	Sequential
	Stereotypical	Egalitarian
4. Individual Differences	Universal	Heritable
	Indp. of general intelligence	Linked to general intelligence
	Indp. of working memory	Limited by working memory capacity

Note. Table adapted from Evans (2008).

Table 3

Predicted Correlations Between the Three Indices and Conflict Problem Accuracy for Each Dual Process Model

	Model	
Index	LIM (De Neys, 2012)	TSM (Pennycook et al., 2015)
Conflict Detection (incorrect conflict RT - nonconflict RT)	Flat, $r = 0$; intercept > 0	positive slope, $r > 0$; intercept ≤ 0
Inhibition (correct conflict RT - incorrect conflict RT)	positive slope, $r > 0$; intercept ≥ 0	No specific prediction
Cognitive Decoupling (correct conflict RT - nonconflict RT)	No specific prediction	negative slope, $r < 0$; intercept ≥ 0

Note. LIM = Logical Intuition Model; TSM = Three-stages Model; RT = Response time.

Table 4

Mean (SD) Accuracy (Proportion of Base-rate Responses) and Response Times (s) in Experiment 1

	Accuracy	Response Time
Problem Type	*M* (*SD*)	*M* (*SD*)
Nonconflict	.90 (.08)	11.6 (3.04)
Conflict	.48 (.30)	
Incorrect Responses		14.3 (4.91)
Correct Responses		14.5 (5.21)

Table 5

Correlations Between Base-Rate Accuracy, CRT Performance, Thinking Dispositions, and Cognitive Ability in Experiment 1

Measure	1	2	3	4	5	6	*M*	*SD*
1. Nonconflict Problem Accuracy	—						.90	.08
2. Conflict Problem Accuracy	.58***	—					.48	.32
3. CRT Performance	.20	.42***	—				.26	.32
4. NFC Composite (Range: 18-90)	.18	.30**	.34**	—			63.00	9.33
5. AOT Composite (Range: 41-246)	.10	.33**	.31**	.29**	—		178.00	19.99
6. SAT Score (out of 2400)	.17	.19	.23*	.15	.28**	—	1847.64	317.48

Note. CRT = Cognitive Reflection Test; NFC = Need for Cognition; AOT = Actively Open-minded Thinking. On the NFC and the AOT, higher numbers represent a greater propensity for engaging in more cognitively-demanding tasks.
*$p < .05$, **$p < .01$, ***$p < .001$

Table 6

Mean (SD) Response Times (s) by Problem Type for Bias-Susceptible and Bias Resistant Groups in Experiment 1

	Bias-Susceptible Group	Bias-Resistant Group
Problem Type	*n* = 44	*n* = 23
Nonconflict		
Incorrect	13.5 (5.52)	15.9 (5.69)
Correct	11.2 (2.65)	11.8 (4.11)
Conflict		
Incorrect	11.9 (2.99)	17.3 (5.27)
Correct	14.4 (5.58)	14.4 (5.01)

Table 7

Mean (SD) Accuracy and Response Times (s) on the Base-rate Neglect and Conditional Reasoning Tasks in Experiment 2

	Base-rate Neglect Task		Conditional Reasoning Task	
	Accuracy	Response Time	Accuracy	Response Time
Problem Type	$n = 102$	$n = 93$	$n = 102$	$n = 102$
Nonconflict	.90 (.11)	13.1 (3.47)	.77 (.12)	8.1 (2.37)
Conflict	.45 (.27)		.40 (.15)	
Incorrect		14.4 (4.00)		7.9 (2.45)
Correct		16.4 (5.07)		8.8 (3.46)

Table 8

Correlations Between Task Accuracy, FOR, and Individual Differences Measures in Experiment 2

Measure	1	2	3	4	5	6	7	8	9	10	11	*M*	*SD*
1. BR Nonconflict Accuracy	—											.90	.11
2. BR Conflict Accuracy	.52***	—										.45	.27
3. BR Nonconflict FOR	.20*	.31**	—									4.69	1.26
4. BR Conflict FOR	-.02	.19†	.89***	—								4.23	1.26
5. CR Nonconflict Accuracy	.10	.05	.02	-.11	—							.77	.12
6. CR Conflict Accuracy	.07	-.05	.14	.17	-.33**	—						.40	.15
7. CR Nonconflict FOR	.07	.12	.64***	.59***	0.19†	.06	—					5.29	1.18
8. CR Conflict FOR	.12	.14	.69***	.62***	.16	.08	.97***	—				5.24	1.18
9. CRT Accuracy	-.02	.12	.09	.07	.06	.26**	.21*	.19*	—			.29	.32
10. NFC Composite	.01	.02	.19†	.30**	.06	.14	.23*	.20*	.21*	—		58.5	11.4
11. SAT Score (out of 2400)	.19†	.29**	.20†	.13	.16	.02	.31**	.30**	.28*	.08	—	1804	282

Note. BR = Base-rate neglect task; CR = Conditional reasoning task; CRT = Cognitive Reflection Test; NFC = Need for Cognition. On the NFC, higher numbers represent a greater propensity for engaging in more cognitively-demanding tasks. For SAT Scores, $N = 91$.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 9

Mean (SD) Accuracy, RT, and FOR Ratings at Pre-Feedback and Post-Feedback in the Base-rate Neglect Task in Experiment 3

		Dependent Measure					
		Accuracy		RT (*s*)		FOR	
Group by Problem Type	*n*	Pre-FB	Post-FB	Pre-FB	Post-FB	Pre-FB	Post-FB
Control	29						
Nonconflict Problems		.90 (.08)	.87 (.11)	**13.5 (1.32)**	**11.5 (2.94)**	4.80 (.32)	4.69 (.92)
Conflict Problems		.46 (.09)	.51 (.24)	**16.8 (1.39)**	**13.2 (3.09)**	4.36 (.31)	4.18 (.82)
False-High FB	34						
Nonconflict Problems		.91 (.08)	.91 (.11)	**14.4 (1.31)**	**12.0 (2.92)**	4.82 (.31)	4.64 (.92)
Conflict Problems		.47 (.09)	.50 (.24)	**16.0 (1.38)**	**12.8 (3.06)**	4.32 (.31)	4.16 (.82)
False-Low FB	31						
Nonconflict Problems		.91 (.08)	.93 (.11)	**14.4 (1.34)**	**10.9 (2.98)**	**4.81 (.32)**	**4.32 (.92)**
Conflict Problems		.45 (.09)	.51 (.24)	**15.9 (1.41)**	**12.4 (3.13)**	4.33 (.32)	4.02 (.82)
True FB	36						
Nonconflict Problems		.90 (.08)	.92 (.11)	**14.5 (1.32)**	**11.3 (2.92)**	4.75 (.32)	4.74 (.91)
Conflict Problems		.46 (.10)	.48 (.24)	**16.0 (1.38)**	**12.5 (3.07)**	4.38 (.31)	4.19 (.82)

Note. FB = Feedback; RT = Response time; FOR = Feeling of Rightness. Bold pairs (Pre-FB & Post-FB) represent a significant change after feedback was presented ($p < .05$).

[a]Adjusted means with pre-FB accuracy covariate value = .68. [b]Adjusted means with pre-FB RT covariate value = 15.2 s. [c]Adjusted means with pre-FB FOR covariate value = 4.57.

Table 10

Correlations Between Base-rate Conflict Problem Accuracy and the Three Indices Before and After Feedback in Experiment 3

		Conflict Detection Index		Inhibition Index		Cognitive Decoupling Index	
Group	*n*	Pre-FB	Post-FB	Pre-FB	Post-FB	Pre-FB	Post-FB
Overall	100	.38***	.43***	-.36***	-.26**	-.15	.04
Control	25	.56**	.50**	-.59**	-.39†	-.32	-.12
False-High FB	24	.32	.38†	-.20	-.32	.01	.01
False-Low FB	22	.25	.47*	-.34	-.19	-.28	.21
True FB	29	.17	.48**	-.15	-.18	-.01	.12

Note. FB = Feedback. Thirty participants were excluded from these analyses for not having any correct ($n = 8$) or incorrect ($n = 22$) conflict responses before or after feedback.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 11

Mean (SD) Accuracy, RT, and FOR Ratings at Pre-Feedback and Post-Feedback in Conditional Reasoning Task in Experiment 3

		Dependent Measure					
		Accuracy[a]		RT (*s*)[b]		FOR[c]	
Group by Problem Type	*n*	Pre-FB	Post-FB	Pre-FB	Post-FB	Pre-FB	Post-FB
Control	29						
Nonconflict Problems		.52 (.09)	.52 (.13)	**9.35 (1.35)**	**5.84 (2.07)**	5.26 (.27)	5.20 (.55)
Conflict Problems		.58 (.09)	.59 (.13)	**8.93 (1.36)**	**6.04 (2.16)**	5.07 (.27)	5.06 (.60)
False-High FB	34						
Nonconflict Problems		.53 (.09)	.51 (.13)	**9.10 (1.35)**	**6.61 (2.07)**	5.28 (.27)	5.40 (.55)
Conflict Problems		.57 (.09)	.56 (.13)	**9.26 (1.36)**	**7.19 (2.15)**	5.03 (.27)	5.21 (.60)
False-Low FB	31						
Nonconflict Problems		.52 (.09)	.51 (.13)	**9.32 (1.35)**	**6.67 (2.07)**	5.16 (.27)	5.06 (.55)
Conflict Problems		.59 (.09)	.56 (.13)	**8.95 (1.36)**	**7.14 (2.16)**	**5.13 (.27)**	**4.88 (.60)**
True FB	36						
Nonconflict Problems		.51 (.09)	.53 (.13)	**9.47 (1.35)**	**6.84 (2.07)**	5.28 (.27)	5.14 (.55)
Conflict Problems		.58 (.09)	.56 (.13)	**8.81 (1.36)**	**6.69 (2.16)**	5.06 (.27)	4.88 (.60)

Note. FB = Feedback; RT = Response time; FOR = Feeling of Rightness. Bold pairs (Pre-FB & Post-FB) represent a significant change after feedback was presented ($p < .05$).

[a]Adjusted means with pre-FB accuracy covariate value = .55. [b]Adjusted means with pre-FB RT covariate value = 9.13 s. [c]Adjusted means with pre-FB FOR covariate value = 5.16.

Table 12

Correlations Between Conditional Reasoning Conflict Problem Accuracy and the Three Indices Before and After Feedback in Experiment 3

		Conflict Detection Index		Inhibition Index		Cognitive Decoupling Index	
Group	*n*	Pre-FB	Post-FB	Pre-FB	Post-FB	Pre-FB	Post-FB
Overall	129	**.23****	**-.01**	-.14	.04	.08	.04
Control	29	.17	-.24	**-.35†**	**.20**	-.35†	-.08
False-High FB	33	.17	.27	-.08	-.24	.15	.03
False-Low FB	31	.28	.04	-.12	.03	.10	.07
True FB	36	**.36**	**-.22**	.03	.30	.26	.10

Note. FB = Feedback. Bold pairs (Pre-FB & Post-FB) show significant ($p < .05$) Fisher's z tests for differences between Pearson correlation coefficients. One participant was excluded from these analyses for not having any incorrect conflict responses before or after feedback.

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Figures

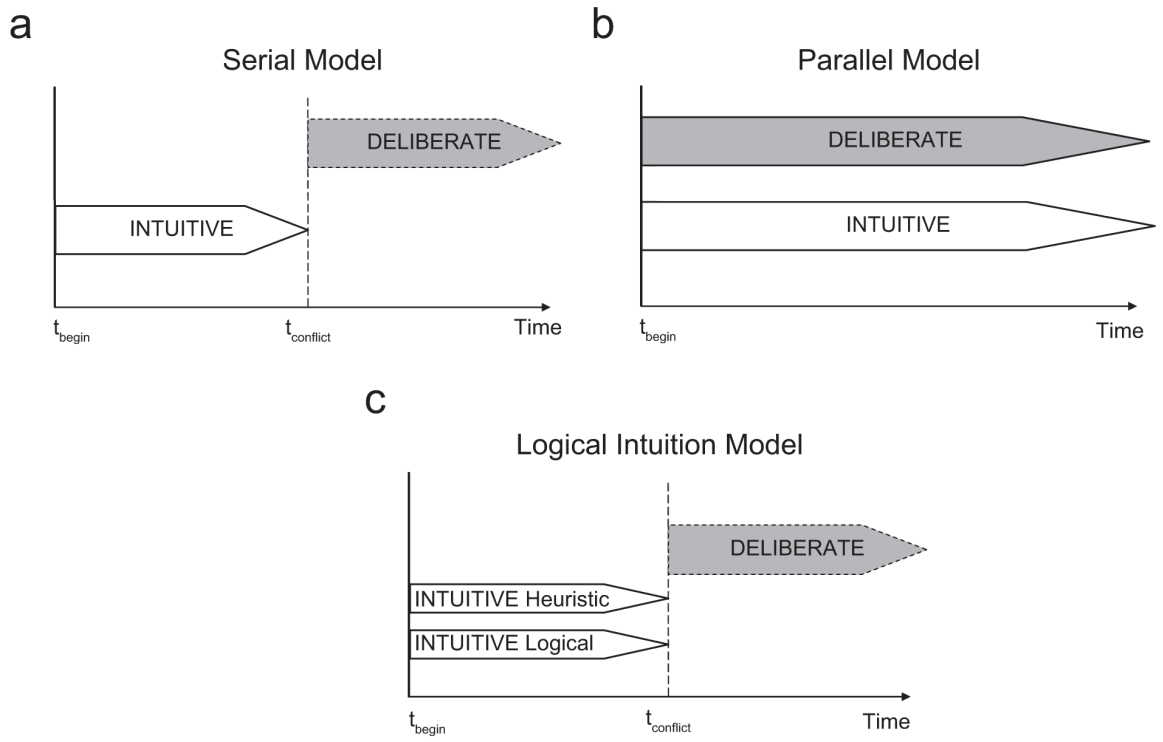

Figure 1. The different theoretical models proposed that illustrates the conflict detection and resolution mechanism within the DPTs framework. In all three representations, the horizontal axis represents time. (A) The serial model; the intuitive system continues to operate until $t_{conflict}$, at which point the deliberate system is optionally activated. (B) The parallel model; the intuitive and deliberate system operate at the same time and $t_{conflict}$ does not have a place to occur. (C) The Logical Intuition Model (LIM); the intuitive system is divided into heuristic and logical intuitions, which operate in parallel, and only if there is a conflict between the two, at $t_{conflict}$, does the deliberate system get optionally activated.[12]

[12] From "Bias and conflict: A case for logical intuitions," by W. De Neys, 2012, *Perspectives on Psychological Science*, p. 34. Copyright 2012 by Wim De Neys. Reprinted with permission.

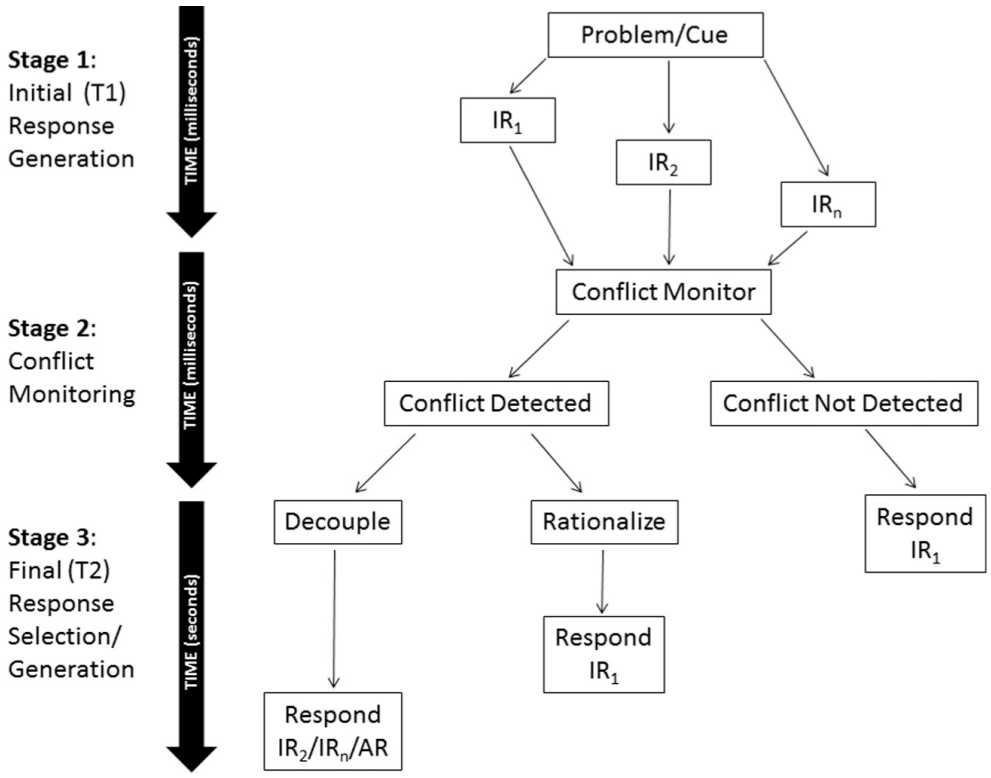

Figure 2. The processes and decisions associated with the Three-stages Model (TSM) utilizing a DPT framework. Initial responses are generated by Type 1 processing and occur within milliseconds in Stage 1. Stage 2 is the conflict monitoring stage and conflict is detected or not detected within milliseconds. Stage 3 represents the engagement of Type 2 processing and whether a person responds with an initial salient response through undetected conflict or rationalization processes, or responds with another initial response as a decoupling process. T1 = Type 1 processing; T2 = Type 2 processing; IR = Initial response; AR = Alternative response.[13]

[13] From "What makes us think? A three-stage dual process model of analytic engagement," by G. Pennycook, J. A. Fugelsang, and D. J. Koehler, 2015, *Cognitive Psychology*, p. 39. Copyright 2015 by Elsevier Inc. Reprinted with permission.

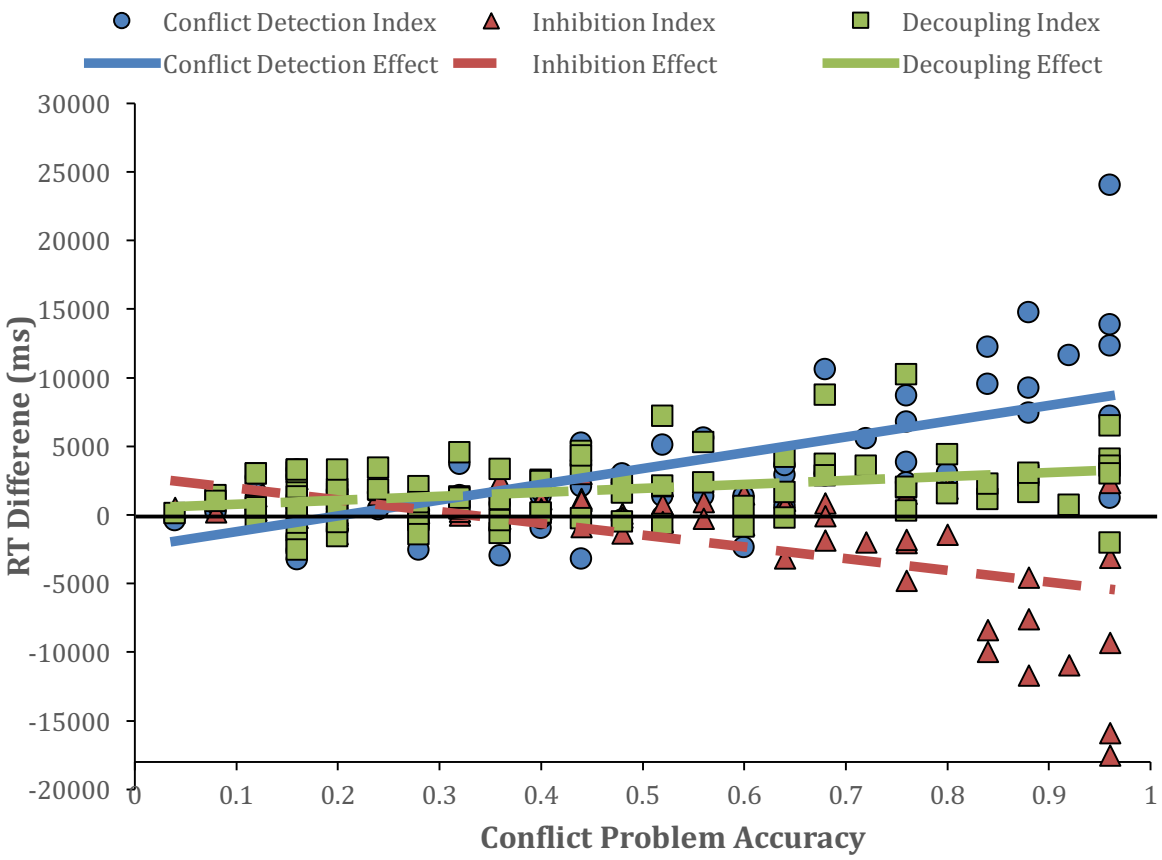

Figure 3. Scatterplot of mean RT time differences and the proportion of correct (base-rate) responses on conflict problems in Experiment 1. Conflict Detection Index refers to the RT difference between incorrect (stereotype) conflict responses and overall nonconflict responses. The Inhibition Index refers to the RT difference between correct and incorrect responses on conflict problems. The Decoupling Index refers to the RT difference between correct conflict responses and overall nonconflict responses. Each unit represents one participant (i.e., one circle, triangle, and square per participant). Lines show regressions of proportion of base-rate responses on RT difference scores.

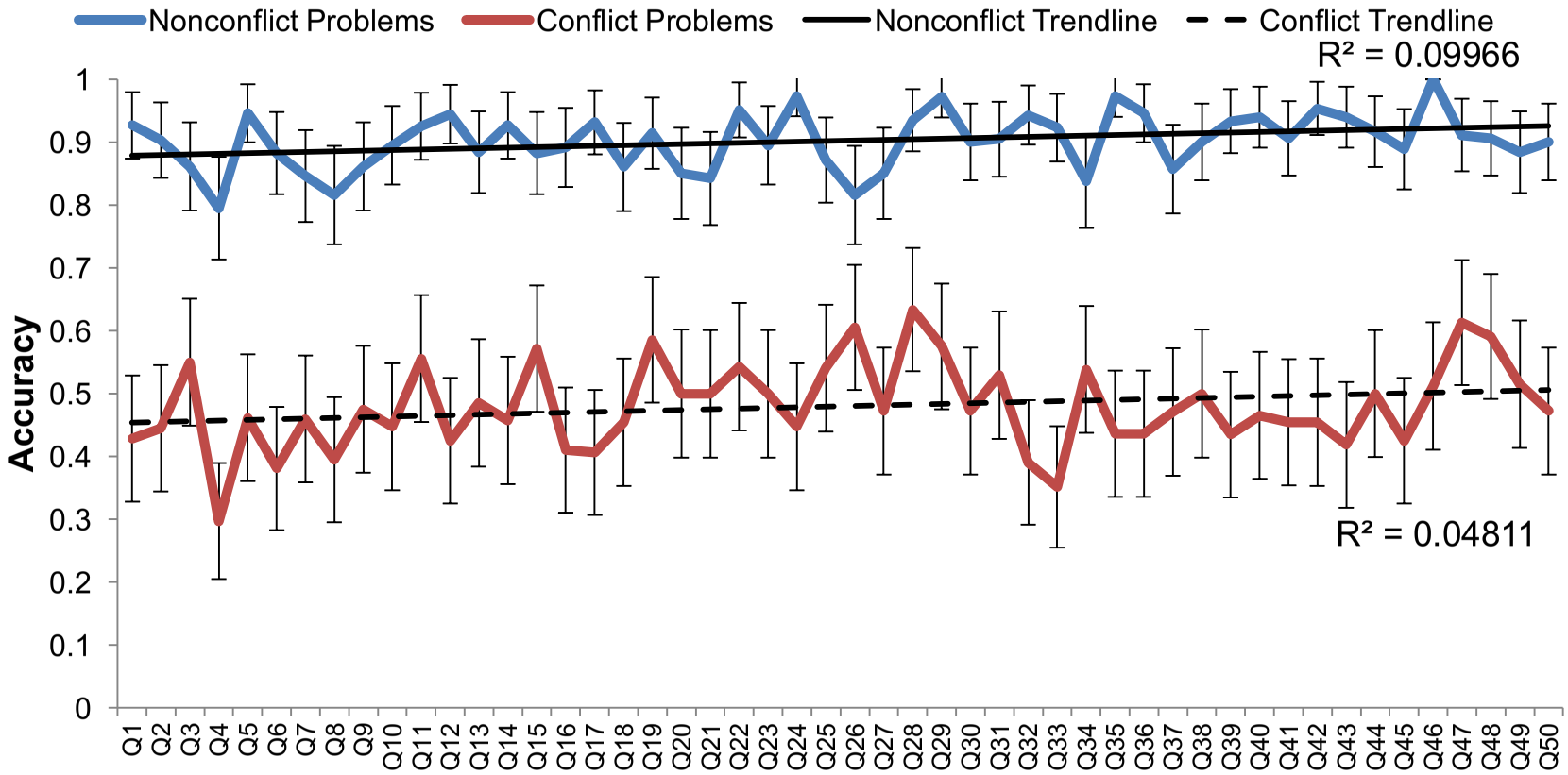

Figure 4. Average performance across all 50 trials separated by problem type. The top line is nonconflict problems with an R^2 value reported for the trend line. The bottom line represents the conflict problems. Each point of both lines represents the average accuracy for those participants that had a specific problem type for a given trial. The lines are independent, as a single participant could not have a nonconflict and a conflict problem for the same trial. Error bars are the standard errors for each mean value per trial.

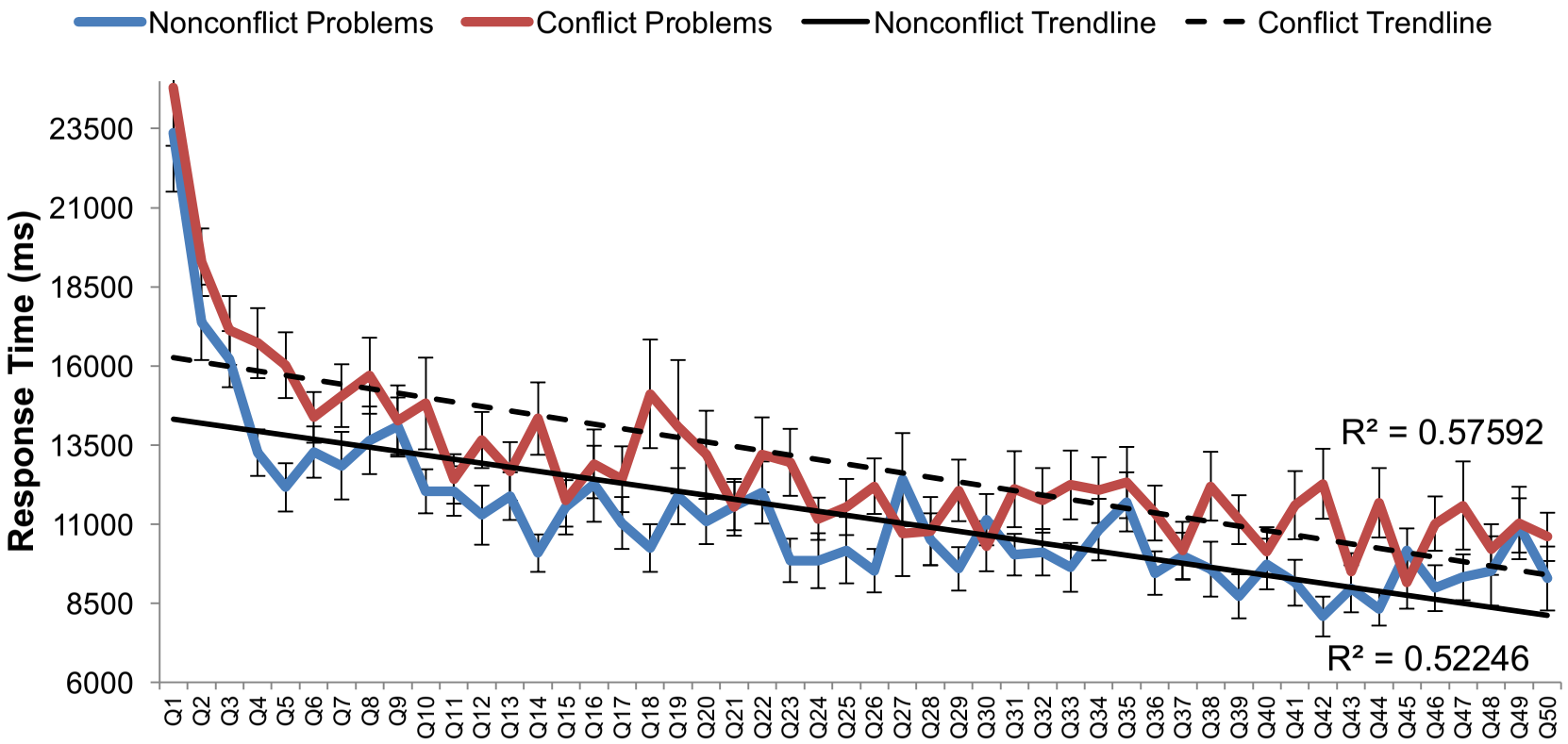

Figure 5. Average response times across all 50 trials separated by problem type. The top line is conflict problems with an R^2 value reported for the trend line. The bottom line represents the nonconflict problems. Each point of both lines represents the average accuracy for those participants that had a specific problem type for a given trial. The lines are independent, as a single participant could not have a nonconflict and a conflict problem for the same trial. Error bars are the standard errors for each mean value per trial.

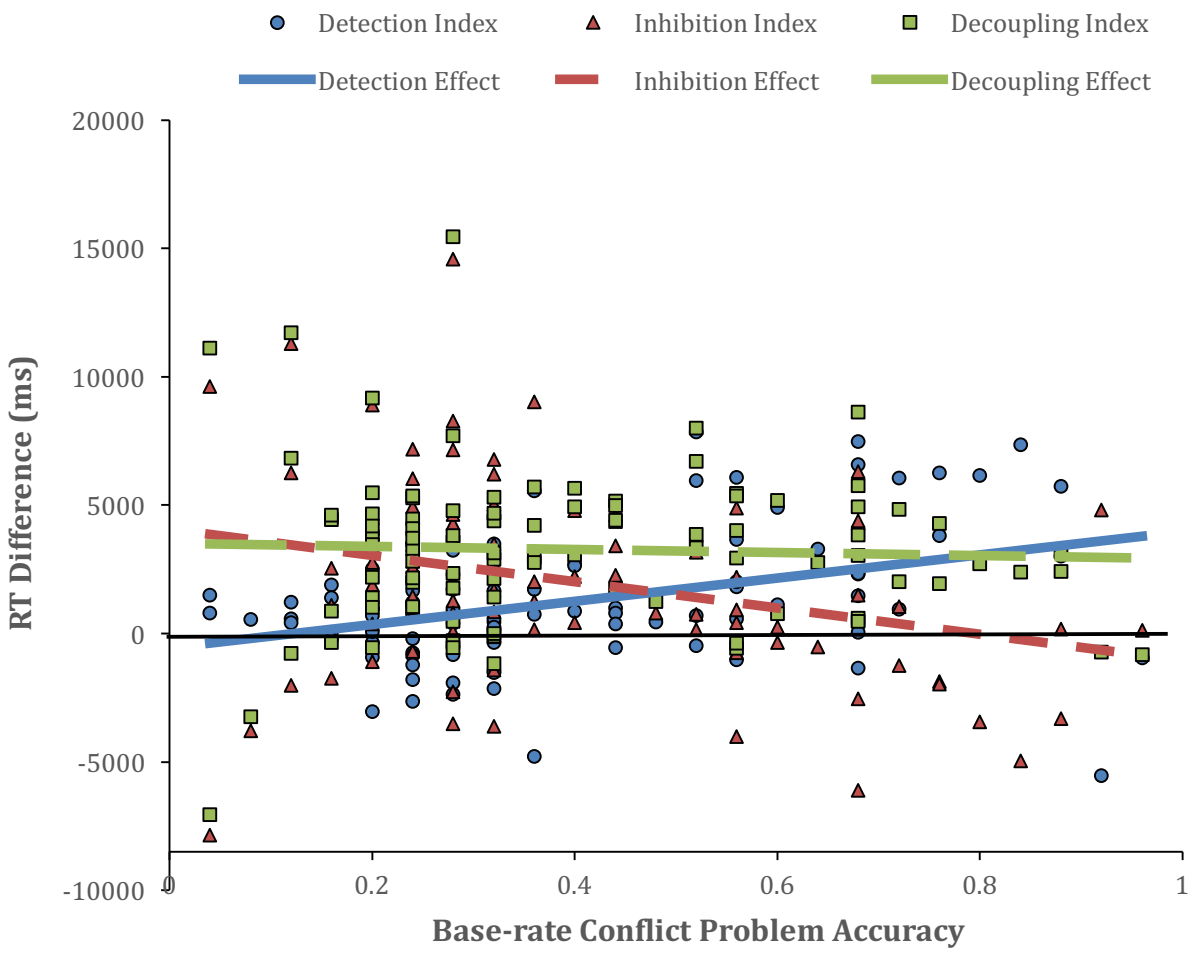

Figure 6. Scatterplot of mean RT time differences and the proportion of correct (base-rate) responses on conflict problems the base-rate neglect task in Experiment 2. Conflict Detection Index refers to the RT difference between incorrect (stereotype) conflict responses and overall nonconflict responses. The Inhibition Index refers to the RT difference between correct and incorrect responses on conflict problems. The Decoupling Index refers to the RT difference between correct conflict responses and overall nonconflict responses. Each unit represents one participant (i.e., one circle, triangle, and square per participant). Lines show regressions of proportion of base-rate responses on RT difference scores.

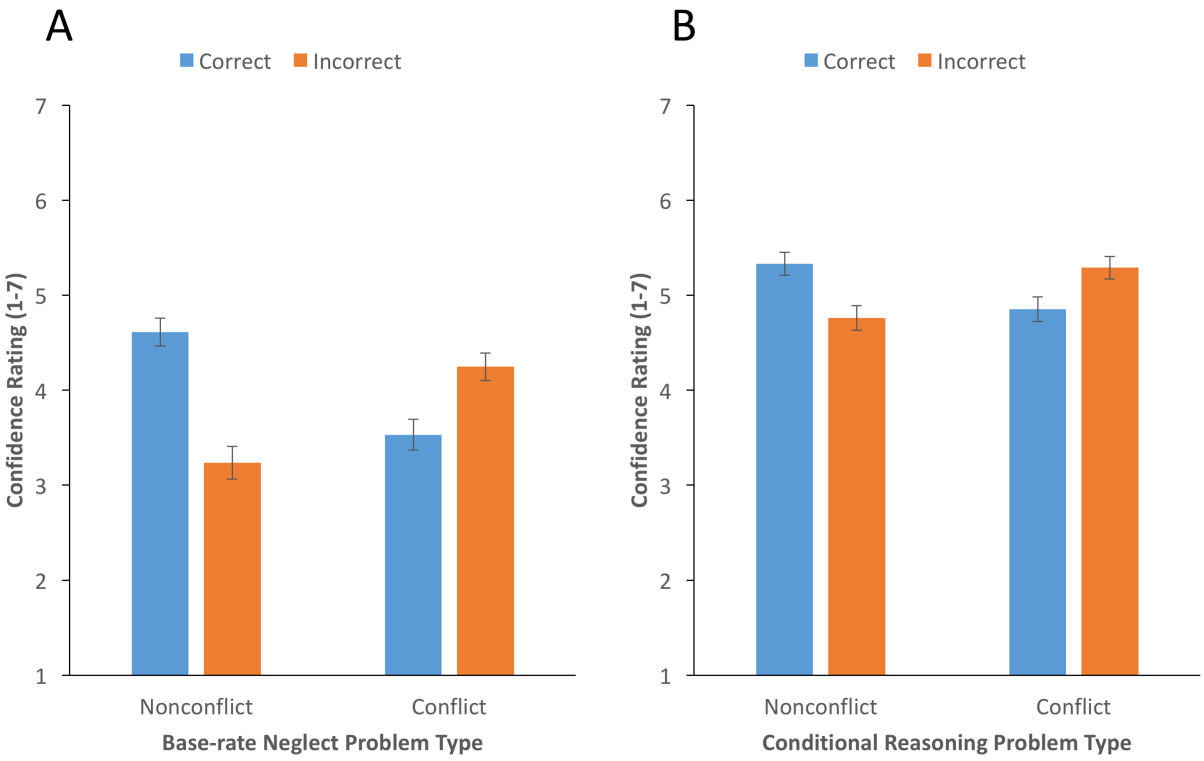

Figure 7. Average confidence/FOR ratings by problem type for both tasks in Experiment 2. Panel A represents the base-rate neglect task. Panel B represents the conditional reasoning task. Error bars represent standard errors.

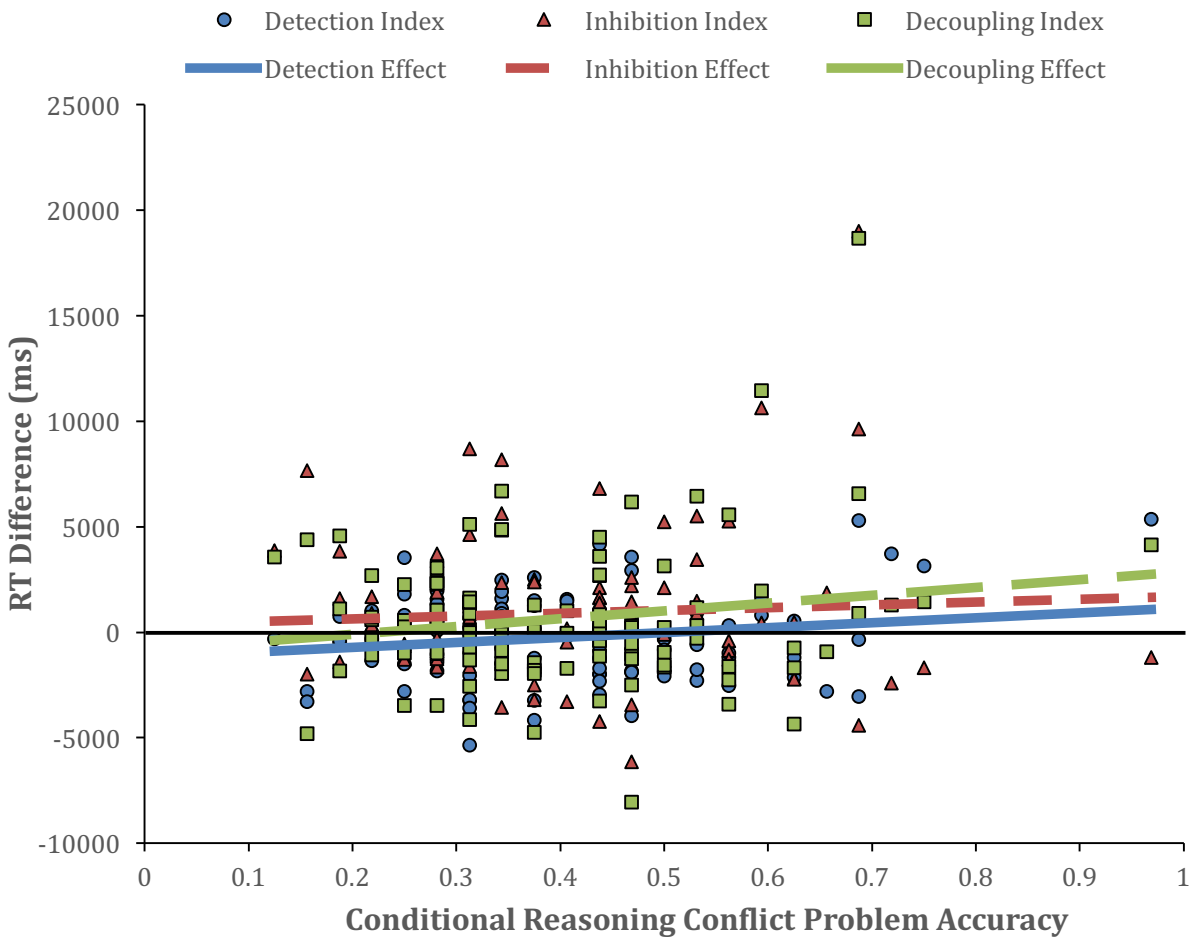

Figure 8. Scatterplot of mean RT time differences and the proportion of correct (YES for valid and NO for invalid) responses on conflict problems the conditional reasoning task in Experiment 2. Conflict Detection Index refers to the RT difference between incorrect (stereotype) conflict responses and overall nonconflict responses. The Inhibition Index refers to the RT difference between correct and incorrect responses on conflict problems. The Decoupling Index refers to the RT difference between correct conflict responses and overall nonconflict responses. Each unit represents one participant (i.e., one circle, triangle, and square per participant). Lines show regressions of proportion of base-rate responses on RT difference scores.

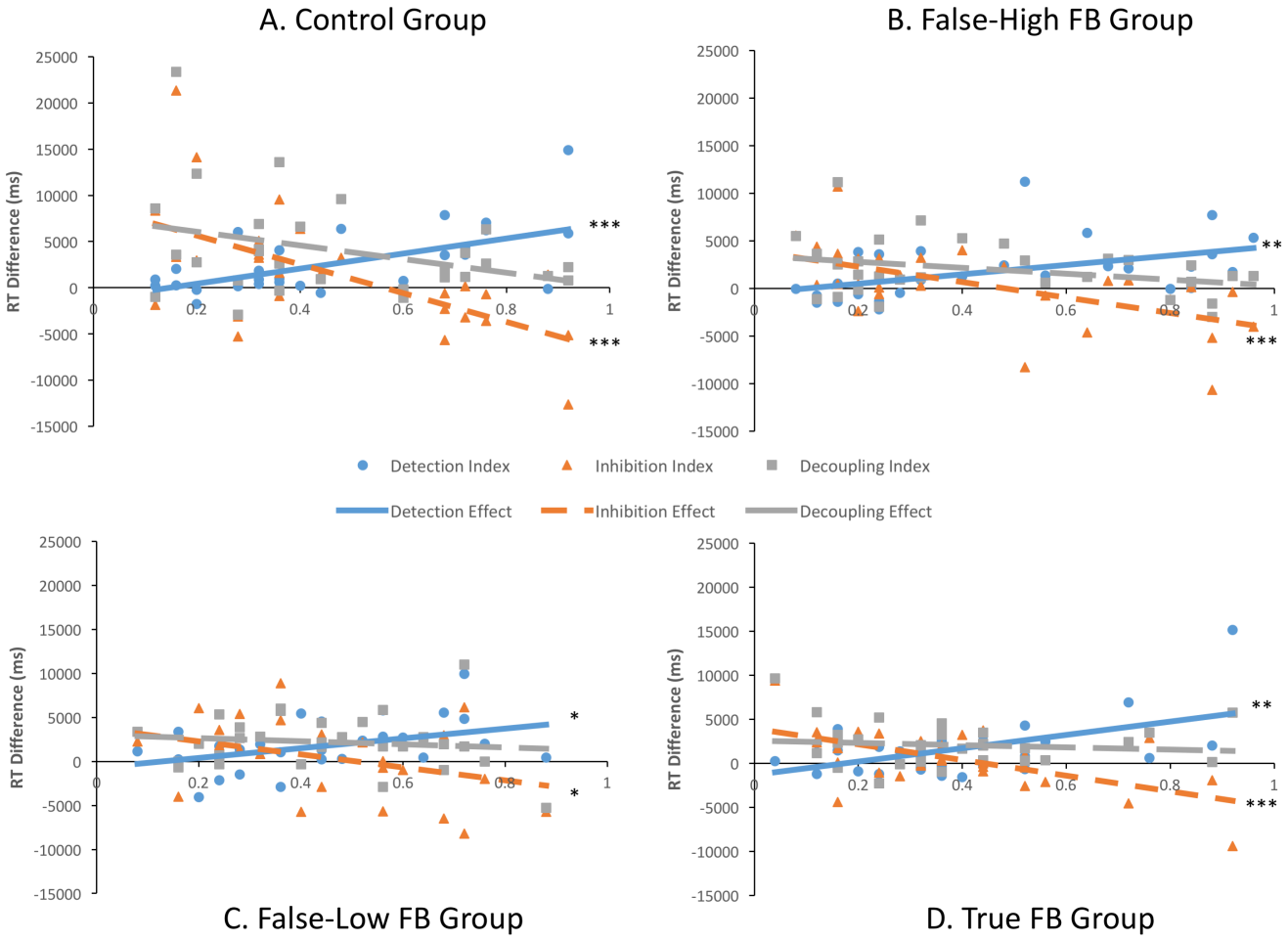

Figure 9. Scatterplots of RT differences (ms) as a function of conflict problem accuracy on the base-rate neglect task separated by group. The scatters and trend lines illustrate the index effects across Feedback (FB). Lines show regressions of proportion of base-rate responses on RT difference scores. *$p < .05$, **$p < .01$, ***$p < .001$

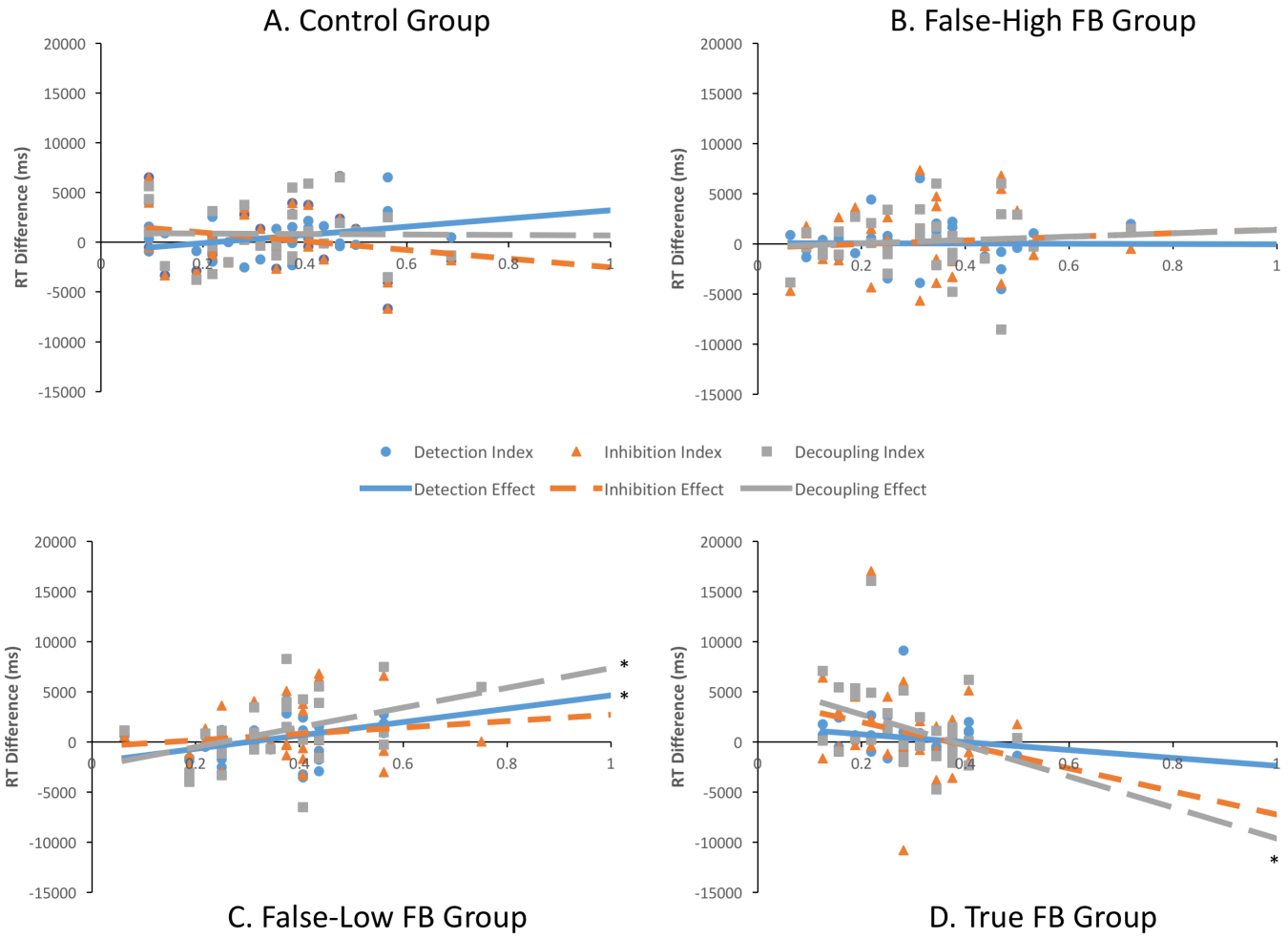

Figure 10. Scatterplots of RT differences (ms) as a function of conflict problem accuracy on conditional reasoning task separated by group. The scatters and trend lines illustrate the index effects across Feedback (FB). Lines show regressions of proportion of correct (Valid/YES and Invalid/NO) responses on RT difference scores. *$p < .05$.

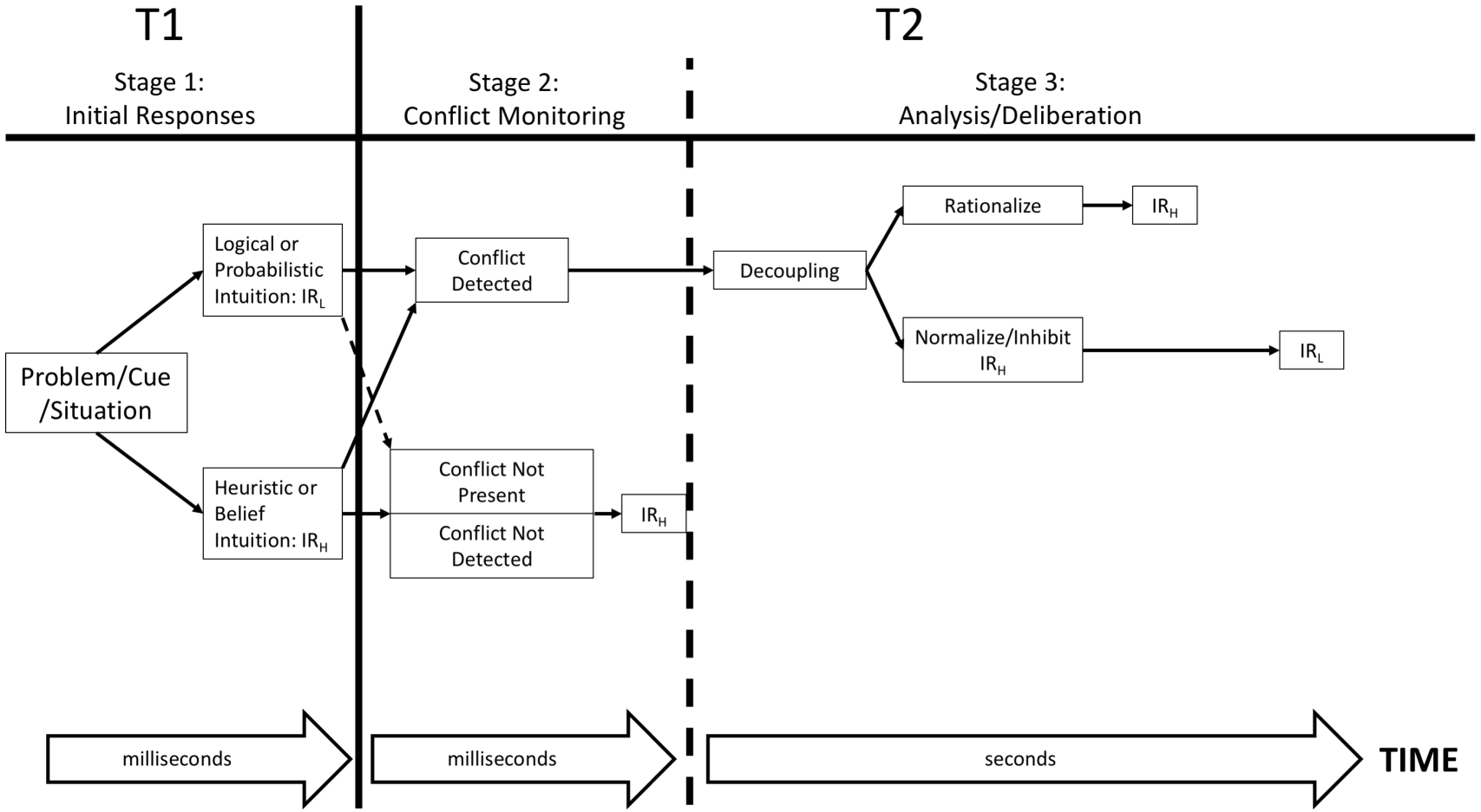

Figure 11. The modified three-stage (TSM) dual process model. T1 = Type 1 processing. T2 = Type 2 processing. IR_H = a heuristic or belief-based initial response. IR_L = a logical or probabilistic intuition initial response. Both intuitive responses are generated by T1 processing, but the dashed line emanating from IR_L to conflict not present/conflict not detected means that it was either not generated or not salient. Final decisions reflect the amount of processing required as show by the length of the arrows.

Appendices

Appendix A: Base-rate Neglect Problems

(Adopted from De Neys, Vartanian, & Goel, 2008)

Nonconflict Problems

1. David is 21 years old. He lives in the rural area and plays soccer. He drinks darker beer in pints and is very witty.

What is more likely?
- a. David is American
- b. David is British

2. Bobby is 23 and works at a nickel refinery. On spare time, Bobby likes to go hunting and enjoys watching hockey at the local pub.

What is more likely?
- a. Bobby is a man
- b. Bobby is a woman

3. Jocelyne is 34. She is married but has no children. She establishes herself in her career and has many political friends.

What is more likely?
- a. Jocelyne is a District Attorney
- b. Jocelyne is a waitress

4. Steve is 23. He works during the school year and brings his laptop to classes. He has no girlfriend and hates going to parties.

What is more likely?
- a. Steve is a computer science major
- b. Steve is a sport management major

5. Rajit is 36. He is a hard worker and loves America. He is religious and spends free time with family when he is not running his corner store.

What is more likely?
- a. Rajit is American-born
- b. Rajit is an immigrant

6. Beverley is 44 years old. She lives in a house outside the city. She is married and has one son and a dog.

What is more likely?
- a. Beverley is a doctor
- b. Beverly is a teacher

7. Jack is 36. He is not married and is somewhat introverted. He likes to spend his free time reading science fiction and writing computer programs.

What is more likely?
- a. Jack is an engineer

b. Jack is a lawyer

8. Jean is 17 years old and works at the neighborhood McDonalds. Jean likes to go out to watch movies, preferably comedies, with friends.

What is more likely?
a. Jean is male
b. Jean is a female

9. Mike is a 26-year-old male. He is married and has two children. He pays taxes on time and drives a mid-size car.

What is more likely?
a. Mike is a catholic
b. Mike is a protestant

10. Sean is a 26-year-old male. He is a college graduate and lives in an apartment. He is single and has no outstanding school loans.

What is more likely?
a. Sean drives a Lexus
b. Sean drives a Ford

11. Wally is 43 and lives in Dover. He is a serious and orderly man. In his free time he studies the history of the British Empire.

What is more likely?
a. Wally is a personal gym trainer
b. Wally is a librarian

12. Jamal is 21 and lives near Brooklyn. Jamal has dreadlocks and drives a convertible. He is 6 ft 7 in and very athletic.

What is more likely?
a. Jamal is a basketball player
b. Jamal is a gymnast

13. John is a lawyer. To ease the stress of work, he spends the day drinking tea and the night drinking beer at his favorite pub.

What is more likely?
a. John is Greek
b. John is English

14. Martine is 26. She is bilingual and reads a lot in her spare time. She is a very fashionable dresser and a great cook.

What is more likely?
a. Martine is American
b. Martine is French

15. Roxy is 20. Her major is Women's Studies and she volunteers at a women's shelter. On the weekends, she teaches a self-defense class.

What is more likely?
a. Roxy is on the debate team

b. Roxy is a cheerleader

16. Catherine is 22. She still lives at home along with her two younger sisters. She spends her summers working as a lifeguard at a summer camp.

What is more likely?
a. Catherine is a history major
b. Catherine is a French major

17. Christopher is 28 years old. He has a girlfriend and shares an apartment with a friend. He likes watching basketball.

What is more likely?
a. Christopher lives in New York
b. Christopher lives in Los Angeles

18. Jo is 23 and is finishing a degree in engineering. On Friday nights, Jo likes to go out cruising with friends while listening to loud music and drinking beer.

What is more likely?
a. Jo is a man
b. Jo is a woman

19. Dieter is very well organized and always on time. So on the weekends he likes to unwind a little bit at one of his favorite techno clubs.

What is more likely?
a. Dieter is German
b. Dieter is Spanish

20. Ian is a 40-year-old male. He lives in the Great Plains and drives a truck. He enjoys listening to country music.

What is more likely?
a. Ian is a farmer
b. Ian is an office worker

21. Antony is 46 and lives in Boston. He is married and has a PhD in mathematics. In his free time, he works on a biography of Sir Isaac Newton.

What is more likely?
a. Antony is a university professor
b. Antony is a carpenter

22. Bree is 23. She is almost 6 feet tall and has a great figure. She likes to wear new clothes and has a personal trainer.

What is more likely?
a. Bree is a model
b. Bree is a maid

23. Russell is 67 and lives in Georgia. He used to work in the oil business and owns a ranch. He believes in the right to bear arms and in traditional marriage values.

What is more likely?
a. Russell is a member of the Green party

b. Russell is a Republican

24. Diego is 25 and has been married for four years. He and his wife have three kids. Diego likes to take a siesta in the afternoon.

What is more likely?
a. Diego is a German
b. Diego is a Mexican

25. Nigel is 42. He is married and has four kids. In his free time, he likes to build model planes and to spend time with his children.

What is more likely?
a. Nigel is a bouncer
b. Nigel is a civil engineer

Conflict Problems

1. Joussef is a 24-year old male who recently married. He has black hair and a dark skin. In his free time he likes to read the Koran.

What is more likely?
a. Joussef is from Finland
b. Joussef is from Iran

2. Luciano is 42. He is quite intelligent. His favorite drink is whiskey and most nights he does not get a lot of sleep.

What is more likely?
a. Luciano is a professional boxer
b. Luciano is a professional poker player

3. Michael is 21 and has a girlfriend. He is self-assured and finds it important to be well-dressed. He is in a debate club and likes sailing.

What is more likely?
a. Michael is a law school student
b. Michael is a musician

4. Andrew is 37 years old. His favorite color is green. In his spare time, he likes to go jogging in his neighborhood and to do work on his car.

What is more likely?
a. Andrew is an office clerk
b. Andrew is a dentist

5. Bob lives in Buffalo. Bob's favorite color is blue. He has a wife Cheryl and a son named Peter. Bob likes to watch television.

What is more likely?
a. Bob is 40 years old
b. Bob is 60 years old

6. Sholan is 26. He lives in LA and likes to wear designer clothes. He acts somewhat stuck-up and plays golf with his father.

What is more likely?

a. Sholan is a plumber
b. Sholan is a stock broker

7. Devon is 20. He grew up in a poor family in a neglected neighborhood in Chicago and didn't finish high school.

What is more likely?

a. Devon is a rapper
b. Devon is a violin player

8. Paul is 34. He lives in a beautiful home in a posh suburb. He is well spoken and very interested in politics. He invests a lot of time in his career.

What is more likely?

a. Paul is a nurse
b. Paul is a doctor

9. Casey is a 36-year-old writer. Casey has two brothers and one sister. Casey likes running and watching a good romantic comedy.

What is more likely?

a. Casey is a man
b. Casey is a woman

10. Marco is 16. He loves to play soccer with his friends, after which they all go out for pizza or to someone's house for homemade pasta.

What is more likely?

a. Marco is Swedish
b. Marco is Italian

11. Stan is 36. He married his college sweetheart after graduating and has two kids. He doesn't drink or smoke but works long hours.

What is more likely?

a. Stan is a dentist
b. Stan is a rock singer

12. Reid is 29. He is quite muscular and is in good shape. He is tanned and has a nice girlfriend.

What is more likely?

a. Reid is a professional surfer
b. Reid is a professional bowler

13. Geoff is from Texas. He is against the president's foreign policy and he throws a party every July 4th.

What is more likely?

a. Geoff is a social worker
b. Geoff is a Marine

14. Angie is 26. She became pregnant at age 16. She's a single mom and lives with her mother. Anita has severe debts.

What is more likely?

a. Angie is an accountant
b. Angie is unemployed

15. Jesse lives in the suburbs, has two kids and drives a brown minivan. In the summer, Jesse likes to spend time at the lake.

What is more likely?

a. Jesse is a man
b. Jesse is a woman

16. Matt is 20 and lives in downtown San Francisco. Matt's favorite food is pasta with meatballs. His parents are living in Seattle.

What is more likely?

a. Matt is a Computer Science major
b. Matt is an English major

17. Trina is 36. When not working, she likes to spend time in the local café, reading poetry and drinking chai lattes. She often changes her hair color.

What is more likely?

a. Trina is a painter
b. Trina is a doctor

18. Kelly is 9. Kelly has a little sister and bugs her all the time. Kelly likes to play with toy trucks and wants to become a famous hockey player.

What is more likely?

a. Kelly is a girl
b. Kelly is a boy

19. Jason is 29 and has lived his whole live in New York. He has green colored eyes and black hair. He drives a light-gray colored car.

What is more likely?

a. Jason is a pool player
b. Jason is a basketball player

20. Lars is 6 ft tall and lives together with his girlfriend in an apartment in the city. He weighs 175 pounds and has short hair.

What is more likely?

a. Lars is Italian
b. Lars is Norwegian

21. Lilly is 37. Her husband is a veterinarian and they have 3 kids. She is committed to her family and always watches the daily cartoon shows with the kids.

What is more likely?

a. Lilly is a kindergarten teacher
b. Lilly is an executive manager

22. Randy is 29. He is quite extraverted and very imaginative. He says he doesn't care about money and is very interested in foreign cultures.

What is more likely?

a. Randy is a theatre actor
b. Randy is a butcher

23. Les is 5 ft 6 tall. Les likes cooking a nice meal and frequently babysits for friends. In high school, Les did well in English but not so well in math.

What is more likely?

a. Les is a man
b. Les is a woman

24. Charlie is 13. Charlie's favorite subject is art. Charlie loves shopping and sleepovers with friends to gossip about other kids at school.

What is more likely?

a. Charlie is a girl
b. Charlie is a boy

25. Fred is 50 years old. He is married and has two daughters. Every morning he reads the newspaper before going to work.

What is more likely?

a. Fred is from New York
b. Fred is from Ohio

Appendix B: Cognitive Reflection Test

(Adopted from Frederick, 2005)

(1) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents

(2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ minutes

(3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____ days

Appendix C: Need for Cognition Scale

(Adapted from Petty, Cacioppo, & Kao, 1984)

For each of the statements below, please indicate whether or not the statement is characteristic of you or of what you believe. For example, if the statement is extremely uncharacteristic of you or of what you believe about yourself (not at all like you) please place a "1" on the line to the left of the statement. If the statement is extremely characteristic of you or of what you believe about yourself (very much like you) please place a "5" on the line to the left of the statement. You should use the following scale as you rate each of the statements below.

1	2	3	4	5
extremely uncharacteristic of me	somewhat uncharacteristic of me	uncertain	somewhat characteristic of me	extremely characteristic of me

1. I prefer complex to simple problems.
2. I like to have the responsibility of handling a situation that requires a lot of thinking.
3. Thinking is not my idea of fun.**
4. I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.**
5. I try to anticipate and avoid situations where there is a likely chance I will have to think in depth about something.**
6. I find satisfaction in deliberating hard and for long hours.
7. I only think as hard as I have to.**
8. I prefer to think about small daily projects to long term ones.**
9. I like tasks that require little thought once I've learned them.**
10. The idea of relying on thought to make my way to the top appeals to me.
11. I really enjoy a task that involves coming up with new solutions to problems.
12. Learning new ways to think doesn't excite me very much.**
13. I prefer my life to be filled with puzzles I must solve.
14. The notion of thinking abstractly is appealing to me.
15. I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.
16. I feel relief rather than satisfaction after completing a task that requires a lot of mental effort.**
17. It's enough for me that something gets the job done; I don't care how or why it works.**
18. I usually end up deliberating about issues even when they do not affect me personally.

Note: ** = reversed coded item.

Appendix D: Actively Open-minded Thinking Scale

(Adapted from Stanovich & West, 1997)

For each of the statements below, mark the alternative that best describes your opinion. There are no right or wrong answers so do not spend too much time deciding on an answer. The first thing that comes to mind is probably the best response.

Response options: 1 – Disagree strongly, 2 – Disagree moderately, 3 – Disagree slightly, 4 – Agree slightly, 5 – Agree moderately, 6 – Agree strongly

1. Even though freedom of speech for all groups is a worthwhile goal, it is unfortunately necessary to restrict the freedom of certain political groups. (Reversed Scored)

2. What beliefs you hold have more to do with your own personal character than the experiences that may have given rise to them. (Reversed Scored)

3. I tend to classify people as either for me or against me. (Reversed Scored)

4. A person should always consider new possibilities.

5. There are two kinds of people in this world: those who are for the truth and those who are against the truth. (Reversed Scored)

6. Changing your mind is a sign of weakness. (Reversed Scored)

7. I believe we should look to our religious authorities for decisions on moral issues. (Reversed Scored)

8. I think there are many wrong ways, but only one right way, to almost anything. (Reversed Scored)

9. It makes me happy and proud when someone famous holds the same beliefs that I do. (Reversed Scored)

10. Difficulties can usually be overcome by thinking about the problem, rather than through waiting for good fortune.

11. There are a number of people I have come to hate because of the things they stand for. (Reversed Scored)

12. Abandoning a previous belief is a sign of strong character.

13. No one can talk me out of something I know is right. (Reversed Scored)

14. Basically, I know everything I need to know about the important things in life. (Reversed Scored)

15. It is important to persevere in your beliefs even when evidence is brought to bear against them. (Reversed Scored)

16. Considering too many different opinions often leads to bad decisions. (Reversed Scored)

17. There are basically two kinds of people in this world, good and bad. (Reversed Scored)

18. I consider myself broad-minded and tolerant of other people's lifestyles.

19. Certain beliefs are just too important to abandon no matter how good a case can be made against them. (Reversed Scored)

20. Most people just don't know what's good for them. (Reversed Scored)

21. It is a noble thing when someone holds the same beliefs as their parents. (Reversed Scored)

22. Coming to decisions quickly is a sign of wisdom. (Reversed Scored)

23. I believe that loyalty to one's ideals and principles is more important than "open-mindedness." (Reversed Scored)

24. Of all the different philosophies which exist in the world there is probably only one which is correct. (Reversed Scored)

25. My beliefs would not have been very different if I had been raised by a different set of parents. (Reversed Scored)

26. If I think longer about a problem I will be more likely to solve it.

27. I believe that the different ideas of right and wrong that people in other societies have may be valid for them.

28. Even if my environment (family, neighborhood, schools) had been different, I probably would have the same religious views. (Reversed Scored)

29. There is nothing wrong with being undecided about many issues.

30. I believe that laws and social policies should change to reflect the needs of a changing world.

31. My blood boils over whenever a person stubbornly refuses to admit he's wrong. (Reversed Scored)

32. I believe that the "new morality" of permissiveness is no morality at all. (Reversed Scored)

33. One should disregard evidence that conflicts with your established beliefs. (Reversed Scored)

34. Someone who attacks my beliefs is not insulting me personally.

35. A group which tolerates too much difference of opinion among its members cannot exist for long. (Reversed Scored)

36. Often, when people criticize me, they don't have their facts straight. (Reversed Scored)

37. Beliefs should always be revised in response to new information or evidence.

38. I think that if people don't know what they believe in by the time they're 25, there's something wrong with them. (Reversed Scored)

39. I believe letting students hear controversial speakers can only confuse and mislead them. (Reversed Scored)

40. Intuition is the best guide in making decisions. (Reversed Scored)

41. People should always take into consideration evidence that goes against their beliefs.

Appendix E: Conditional Reasoning Problems

(Adopted from Thompson, 1994)

NS Type	Pragmatic Context	Temporal Sequence	Statement
+N+S	Causal	A	If butter is heated, then it melts
+N+S	Causal	B	If butter melts, then it has been heated.
+N+S	Causal	A	If water is heated to 100 degrees centigrade, then it boils.
+N+S	Causal	B	If water boils, then it has been heated to 100 degrees centigrade.
+N+S	Definition	A	If an animal is warm-blooded, then it is a mammal.
+N+S	Definition	B	If an animal is a mammal, then it is warm-blooded.
+N+S	Definition	A	If a person is someone's mother's mother, then she is that person's maternal grandmother.
+N+S	Definition	B	If a person is someone's maternal grandmother, then she is that person's mother's mother.
-N+S	Causal	A	If the car is out of gas, then it stalls.
-N+S	Causal	B	If the car stalls, then it is out of gas.
-N+S	Causal	A	If the dog tracks mud on the floor, then the floor is dirty.
-N+S	Causal	B	If the floor is dirty, then the dog has tracked mud on it.
-N+S	Definition	A	If an animal is a robin, then it is a bird.
-N+S	Definition	B	If an animal is a bird, then it is a robin
-N+S	Definition	A	If a card has a jack on it, then it is a face card.
-N+S	Definition	B	If a card is a face card, then it has a jack on it.
+N-S	Causal	A	If the T.V. is plugged in, then it works.
+N-S	Causal	B	If the T.V. works, then it is plugged in.
+N-S	Causal	A	If the car has gas in it, then it runs.
+N-S	Causal	B	If the car runs, then it has gas in it.
+N-S	Definition	A	If a plant has roots, then it is a tree.
+N-S	Definition	B	If a plant is a tree, then it has roots.
+N-S	Definition	A	If a figure has 4 sides, then it is a square.
+N-S	Definition	B	If a figure is a square, then it has four sides.
-N-S	Causal	A	If the weather conditions are bad, then the plane will crash.
-N-S	Causal	B	If the plan has crashed, then the weather conditions have been bad.
-N-S	Causal	A	If a person eats toffee, then they get cavities.
-N-S	Causal	B	If a person gets cavities, then they eat toffee.
-N-S	Definition	A	If a piece of fruit is red, then it is an apple.
-N-S	Definition	B	If a piece of fruit is an apple, then it is red.
-N-S	Definition	A	If a piece of furniture is made of wood, then it is chair.
-N-S	Definition	B	If a piece of furniture is a chair, then it is made of wood.