Dual-System Theories of Decision Making: Analytic Approaches and Empirical Tests

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Aleksandr Sinayev

Graduate Program in Psychology

The Ohio State University

2016

Dissertation Committee:

Ellen Peters, Adviser

Michael DeKay

Trisha Van Zandt
Abstract

Dual-system models are popular in the study of decision making. They have faced criticisms, especially for being vague and lacking specific predictions. In order to address these criticisms, three categories of dual-system models are reviewed: parallel-competitive (in which intuitive, system 1, and deliberative, system 2, processing happen at the same time and both influence the response), default-interventionist (in which system 1 executes first and then system 2 may or may not override system 1), and interactive (in which both systems process information at the same time, but they are allowed to influence each other in complex back-and-forth interactions). Tests are conducted of the former two categories.

Default-interventionist dual-system models predict that individual differences in reflectiveness should be associated with less biased decision making. The Cognitive Reflection Test (CRT) is thought to measure monitoring of system 1 intuitions such that, if cognitive reflection is high enough, intuitive errors will be detected and the problem will be solved. However, CRT items also require numeric ability to be answered correctly and it is unclear how much numeric ability vs. cognitive reflection contribute to better decision making. In two studies, CRT responses were used to calculate Cognitive Reflection and numeric ability; a numeracy scale was also administered. Numeric ability, measured with the CRT or the numeracy scale, accounted for the CRT's ability to predict
more normative decisions (a subscale of decision-making competence, incentivized measures of impatient and risk-averse choices, and self-reported financial outcomes); Cognitive Reflection contributed no independent predictive power. Results were similar whether the two abilities were modeled (Study 1) or calculated using proportions (Studies 1 and 2). These findings demonstrated that correlations of decision performance with the CRT are insufficient evidence to implicate overriding intuitions, and, therefore, to support default-interventionist theories in the decision-making biases and outcomes we examined. Numeric ability appeared to be the key mechanism instead.

After failing to find evidence for a default-interventionist model of decision making, a mathematical parallel-competitive model, first introduced by Mukherjee (2010), was reviewed and tested. This model was expanded in order to improve its fit to choice data. An experiment was conducted, in which participants made risky choices. Half of the participants made choices under heavy cognitive load, and the other half under no cognitive load. The model’s key parameter indicating the use of system 1 vs. system 2 changed in the predicted way, but the model failed to account for important features in the data. The model’s assumed system 1 process was changed to be sensitive to probability, rather than outcome, after which the model was able to account for qualitative patterns of data more effectively and fit the data better, suggesting this is a better representation of system 1.

In the final chapter, I conclude that some dual-system models are specific enough to be mathematically identifiable and to make a number of very specific predictions. I then
discuss the possibility of generalizing the analytic approaches I employed to other
decision problems and the limitations of these approaches.
Acknowledgments

I would like to thank my adviser, Ellen Peters, for encouraging me throughout my studies and for always being in my corner. I have always left meetings feeling like no challenge is insurmountable and have always felt she is on my side.

I would like to thank my whole committee for being patient with me as deadlines were moved and for providing endless crucial suggestions.

I would like to thank my parents and brother for providing an emotional safety net during trying times, providing direction during times of confusion and for always believing I am worthy.
Vita

May 2007 .................................................. Brooklyn Technical High School

May 2011 .................................................. B.A. Psychology, Binghamton University

May 2011 .................................................. B.A. Mathematics, Binghamton University

May 2013 .................................................. M.A. Psychology, The Ohio State University

Sept 2011 – Aug 2012 ......................... Fellow, Department of Psychology, The Ohio State University

Sept 2012 - to present ...................... Graduate Teaching Associate, Department of Psychology, The Ohio State University

Sept 2013 - to Aug 2014 ...................... Graduate Research Assistant, Department of Psychology, The Ohio State University

Publications


Fields of Study

Major Field: Psychology
Table of Contents

Abstract ................................................................................................................................. ii

Acknowledgments .............................................................................................................. v

Vita ........................................................................................................................................ vi

List of Tables ..................................................................................................................... xii

List of Figures ................................................................................................................... xiii

Chapter 1: Introduction .................................................................................................... 1

Significance of dual-system models .................................................................................. 2

Review of dual-system models in reasoning and decision making ................................. 3

Types of models ................................................................................................................ 4

Importance of working memory in dual-system models ................................................. 25

Alternatives to and criticisms of dual-system models ..................................................... 26

Review of single continuum models .............................................................................. 27

Review of multiple systems models in relation to dual-system models ...................... 33

Other criticisms of dual-system models ........................................................................... 39

Summary ............................................................................................................................. 40

Transition to Chapter 2 ...................................................................................................... 41

viii
Chapter 2: Cognitive Reflection vs Numeric ability in predicting decision biases and outcomes (Sinayev & Peters, 2015) ................................................................. 43

The Cognitive Reflection Hypothesis ............................................................... 43

Dual-System Explanation .................................................................................. 45

The Numeracy Hypothesis ................................................................................ 47

Modeling Cognitive Reflection and Numeric Ability ....................................... 48

Do Cognitive Reflection and Numeracy Both Predict Good Decision Making? 50

Study 1 ............................................................................................................... 52

Method ............................................................................................................ 53

Results ............................................................................................................. 58

Discussion ....................................................................................................... 66

Study 2 ............................................................................................................... 68

Method ............................................................................................................ 70

Results ............................................................................................................. 76

Discussion ....................................................................................................... 83

General Discussion .......................................................................................... 85

Transition to Chapter 3 ..................................................................................... 93

Chapter 3: Experimental test of dual-system predictions in Risky Choice ......... 96

Introduction ..................................................................................................... 96
Mukherjee’s (2010) Dual-system Model ................................................................. 98
Risky-choice findings relevant to working-memory-load manipulations .......... 111
Method ............................................................................................................ 113
Sample .......................................................................................................... 113
Procedure ...................................................................................................... 114
Design .......................................................................................................... 116
Results .......................................................................................................... 117
MDSM model fits .......................................................................................... 122
Alternative model fits ................................................................................... 131
Is the δ parameter doing all the work? .......................................................... 137
Discussion ...................................................................................................... 137

Chapter 4: GENERAL DISCUSSION .................................................................... 142

Evidence from Chapters 2 and 3 for and against Dual-System Theories ............ 143

Generalizations to Other Decision Tasks ...................................................... 145

Difficulties in Applying Computational Methods to Interactive Models ........ 146

Relation to Affect and Cognitive Heuristics ................................................... 147

Limitations of Repeated Measures Methodologies ........................................ 148

Aggregation Errors ....................................................................................... 150

Additional Limitations .................................................................................. 151
List of Tables

Table 1. Summary of the dual-system models reviewed .................................................. 5
Table 2. Common criticisms of dual-system theories .................................................... 27
Table 3. Coding of Possible Responses ....................................................................... 56
Table 4. Intuitive and Correct Responses for CRT Items Used in Study 1 ................. 60
Table 5. Correlations of the Measures in Study 1 ......................................................... 61
Table 6. Model difficulty parameters (standard errors) for each CRT item ............... 63
Table 7. Regression Analyses - Consistency in Risk Perception and CRT ............... 65
Table 8. Financial Outcomes ....................................................................................... 74
Table 9. Correlations of independent measures among themselves and with DVs .... 78
Table 10. Stepwise regression results predicting decision-bias and financial-outcome composites ........................................................................................................... 82
Table 11. Robust regression results for Chapter 1 Study 1 ......................................... 170
Table 12. Robust regression results for Chapter 1 Study 2 ......................................... 171
List of Figures

Figure 1. A stylized and simplified graphical representation of the archetypal parallel-competitive model................................................................................................................................. 7

Figure 2. A graphical representation of a stylized, simplified default-interventionist model. The dashed arrow indicates that system 2 may or may not be activated and inhibit system 1’s intuition. In Kahneman’s theory, the figure would have additional arrows. .. 14

Figure 3. Mukherjee’s (2010) graphic representation of the dual-system model......... 99

Figure 4. Mean proportions and 95% confidence intervals by condition. ............... 119

Figure 5. A graphical depiction of the Bayesian MDSM.. ........................................ 123

Figure 6. Priors and posteriors on key MDSM parameters. ........................................ 127

Figure 7. Mean and MDSM estimated proportion choosing the certain option. .......... 129

Figure 8. Mean and MDSM estimated proportion choosing a higher likelihood gamble. ........................................................................................................................................................................... 130

Figure 10. Mean and modified model estimated proportion choosing the certain option. ........................................................................................................................................................................... 135

Figure 11. Mean and modified model estimated proportion choosing a higher likelihood gamble........................................................................................................................................................................... 136

Figure 12. MDSM fits without delta to proportions choosing certain option............ 175

Figure 13. MDSM fits without delta to proportions choosing higher likelihood gamble176
Figure 14. Expected value model fits to proportion choosing certain option................. 181
Figure 15. Expected value model fits to proportions choosing higher likelihood gamble

........................................................................................................................................ 182
CHAPTER 1: INTRODUCTION

A question of central interest in psychology is whether one system of thought (a set of thought processes that share some characteristics) is responsible for a psychological phenomenon, whether it is necessary to postulate an additional, distinct system of thought to explain this phenomenon, or whether more than two distinct systems are necessary. Recently, this idea has become popular in the study of decision making, but it is by no means new.

In one form or another, this idea has permeated most branches of psychology and goes back to the foundation of the discipline. William James (1890/1950) explicitly postulated two types of thought: associative and true reasoning, with the former being useful for creative pursuits and the latter being useful for analytic pursuits. Behaviorism, on the other hand, postulated that all human mental activity can be accounted for by non-deliberative processes, like conditioning, and some behaviorists went as far as to say that deliberative processes simply occur to create post-hoc explanations (Watson, 1913). This reasoning suggests that this second deliberative system of thought is unnecessary and will only confuse our understanding of the human mind. More recently, dual-system models have regained popularity, first in the study of memory, then in explaining biased cognition in fields of reasoning and decision making.
Dual-system models in decision making and criticisms and tests of these models are the emphasis of the present work. This chapter suggests that dual-system models in psychology can be grouped by their functional attributes. This chapter also discusses criticisms of dual-system models and possible alternatives. The chapters that follow attempt to address some of the criticisms of dual-system models by providing empirical tests for two of the three types of dual-system models discussed in this chapter.

**Significance of dual-system models**

Dual-system models have a long history, but they have gained recent prominence in cognitive psychology (e.g., Sloman, 1996) and social psychology (Chaiken & Trope, 1999) and have motivated many experimental studies. Not all researchers agree that these studies have adequately tested dual-system predictions (e.g., Keren & Schul, 2009; Gigerenzer, 2010; Kruglanski & Gigerenzer, 2011; reviewed in more depth below), but it is clear that many studies have used this framework based on the numbers of citations of some of these papers. For example, Epstein’s (1994) Cognitive-Experiential Self-Theory, reviewed here, had 2688 citations on Google Scholar as of 7/22/15; Sloman’s (1996) dual system account had 2718.

Dual-system models have gained popularity in decision-making research in particular. Currently, they are used as an overarching theme to explain important findings in the field (Kahneman, 2011; Thaler & Sunstein, 2008). However, dual-system models are grouped together as if they were identical, despite the fact that many dual-system theorists do not agree with each other. For example, some authors emphasize the role of
affect (e.g., Epstein, 1994), while others largely ignore it (e.g., Sloman, 1996), a distinction important in decision-making research (Slovic et al., 2004). Moreover, dual-system models in the domain of decision making have only been tested in a handful of paradigms and with mixed success. The present research attempts to fill this gap in part by reviewing research on dual-system models in a systematic way and using analytic techniques (IRTrees in Chapter 2 and Bayesian modeling in Chapter 3) that are novel for research in dual-system models in decision making.

**Review of dual-system models in reasoning and decision making**

This section is intended to provide an overview of the various dual-system models relevant to decision making. The generic use of the term “system” in the present paper is motivated by the fact that most of the theories reviewed have used this term. For simplicity’s sake, the term “dual-system” will be used interchangeably with “dual modes of thought” and “dual-process.” Moreover, the use of the same term between different types of theories is for continuity, and not intended to imply that all the theories are talking about the same systems (instead, it will be made clear that dual-system theorists do not always agree).

A complete review of all dual-system models is outside the scope of the present work (e.g., Sherman, Gawronski & Trope, 2014 take an entire volume to provide an update of such theories in social psychology since 1999). An attempt is made to group dual-system models by several features and to describe exemplars of models of each type most relevant to decision making. In particular, the present review emphasizes default-
interventionist and parallel-competitive models (defined below) because Chapters 2 and 3, respectively, provide analytic approaches to testing these types of models. Three other types of models (interactive, single continuum, and multiple system models) are also briefly discussed to provide context.

**Types of models.**

One meaningful method of classifying dual-system theories is through functional distinctions (i.e., the mechanism proposed in the model, rather than, for example, the domain to which it is applied). All dual-system theories posit a psychological (i.e., latent) explanation for a set of behavioral responses in which two systems with different properties produce output, and output from one or both of these systems results in behavior. The exact mechanics of how behaviors are derived from these outputs, however, as well as the properties of the systems, and consequently their outputs, differ between theories. Specifically, three types of theories will be discussed: parallel-competitive theories (which assume processing in the two systems happens separately and simultaneously), default-interventionist theories (which assume that one system provides defaults and the other sometimes overrides them) and interactionist theories (which allow complex back-and-forth interactions between the two systems). The models are described in some detail below and summarized in Table 1.
Table 1. Summary of the dual-system models reviewed.

<table>
<thead>
<tr>
<th>Name and author(s)</th>
<th>System 1 Key Properties</th>
<th>System 2 Key Properties</th>
<th>Mechanics</th>
<th>Distinct Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Experiential Self Theory (Epstein)</td>
<td>Emotional, unconscious</td>
<td>Rational, conscious</td>
<td>Parallel-competitive</td>
<td>Experiential system influences rational system. Leaves open possibility of other interactions.</td>
</tr>
<tr>
<td>Sloman’s Dual-system Theory</td>
<td>Associative</td>
<td>Rule-based</td>
<td>Parallel-competitive</td>
<td>Both can be conscious</td>
</tr>
<tr>
<td>Mukherjee’s Dual-system Model</td>
<td>Affective</td>
<td>Deliberative</td>
<td>Parallel-Competitive</td>
<td>Simplified, mathematical model</td>
</tr>
<tr>
<td>Kahneman’s Dual-system Theory</td>
<td>Fast, automatic</td>
<td>Slow, controlled</td>
<td>Default-Interventionist</td>
<td>Both can be based on language</td>
</tr>
<tr>
<td>Stanovich’s Dual-system Theory</td>
<td>No Cognitive Decoupling</td>
<td>Involves Cognitive Decoupling</td>
<td>Default-Interventionist</td>
<td>Allows many sub-systems for system 1</td>
</tr>
<tr>
<td>Evans’ Dual-system Theory</td>
<td>Does not require working memory</td>
<td>Requires working memory</td>
<td>Default-Interventionist</td>
<td>Repeated proposals from system 1</td>
</tr>
</tbody>
</table>

To begin, Evans and Stanovich (2013) defined parallel-competitive and default-interventionist types of designs as follows:

S. A. Sloman (1996; Barbey & Sloman, 2007)... proposed an architecture that has a parallel-competitive form. That is, Sloman’s theories and others of similar structure (e.g., Smith & DeCoster, 2000) assume that Type 1 and 2 processing proceed in parallel, each having their say with conflict resolved if necessary. In
contrast, our own theories (in common with others, most notably that of
Kahneman & Frederick, 2002; see also, Kahneman, 2011) are default-
interventionist in structure (a term originally coined by Evans, 2007). Default-
interventionist theories assume that fast Type 1 processing generates intuitive
default responses on which subsequent reflective Type 2 processing may or may
not intervene. (pp. 227)

As will become clear, many dual-system models do not fall squarely into one of these
two types. However, these two types of dual-system models appear to be sufficient to
capture the main features of the dual-system models most often applied to reasoning and
decision making.

**Parallel-Competitive Models**

The defining features of parallel-competitive models are that systems 1 and 2 operate
continuously (they do not stop), simultaneously (rather than sequentially), and compete
for the overt response. As with other types of dual-system models, parallel-competitive
models differ regarding the defining characteristics of either system (which
characteristics are necessary and sufficient for calling something a system 1 phenomenon,
or even whether such characteristics exist), but they usually agree upon the same
correlated characteristics. A simple version of a parallel-competitive model is shown in
Figure 1. Many theorists add features to the pure parallel-competitive model (e.g., system
1 resolves faster than system 2), but this simple version has the advantage of being easy
to quantify. It is also a general version of Mukherjee’s (2010) model used in Chapter 3 to
mathematically model the decision-making process in a risky choice paradigm. In this section, I review the main parallel-competitive models that have been applied to judgment and decision research.

Figure 1. A stylized and simplified graphical representation of the archetypal parallel-competitive model.

Epstein (1994) described a parallel-competitive model called Cognitive-Experiential Self-Theory (CEST) in which system 1 (the Experiential System) has a heavy emphasis on emotion, whereas system 2 (the Cognitive System) is rational and logical. In CEST, both system 1 and system 2 are always active and always competing for responses. It appears that one defining distinction is that system 1 is unconscious and emotional and system 2 is conscious and logical, making a direct connection to Freud’s psychodynamic theory.

Epstein also subscribes some of the usual correlates of each system including that system 1 is faster, more concrete, and reliant on associations, whereas system 2 is slower, more abstract, and reliant on logical connections.

Epstein (1994) and Epstein and Pacini (1999) also suggested there is an order of operations in CEST. Specifically, they suggest that in response to an event, system 1 will
act first, looking for associations and their attached emotional valences, which then shape the further mental events including conscious and unconscious thoughts and actions. If the retrieved associations are positive, the thoughts and actions will be guided by a desire to reproduce the retrieved events. Otherwise, they will be guided by a desire to avoid them. The valence of associations is determined with regards to how they fit with humans’ four basic needs: 1) to seek pleasure and avoid pain, 2) to maintain a coherent conceptual system, 3) to have social contact, and 4) to maintain self-esteem. Epstein and Pacini suggest that to assign valence, all aspects of subjective reality are rated with regards to their ability to fulfill these needs (these ratings, in turn, are derived from previous experience). Although they emphasize system 1’s effects on system 2, Epstein and Pacini, in principle, leave open any sort of interaction between the two systems. “Each can influence the other in terms of content,” Thus, Figure 1 would technically need to have a bidirectional arrow between system 1 and system 2 to account for their theory. Because Epstein & Pacini do not describe how or when system 2 might influence system 1, I contend the interactive nature of this model is not central and classify their model as parallel-competitive as opposed to interactionist.

Although CEST has been classified as a parallel-competitive model (Evans, 2008), this classification may not be fully consistent with some of its features. First, though CEST maintains that the two systems work in parallel (both of them influence the person at decision time), CEST’s other claims are consistent with consecutive processing (system 1 is able to complete processing before system 2 processing begins), which is more
consistent with a default-interventionist model as discussed in the next section. Second, Epstein leaves the two systems open for interaction (consistent with interactionist models discussed later), but only mentions one specific way in which they interact (i.e., system 1 attempts to make any deliberative processing consistent with it).

A second popular parallel-competitive theory is Sloman’s (1996) dual-system theory. Sloman calls his system 1 the Associative System and system 2 the Rule-based System. Sloman attempted to build his theory based on neurocomputational work popular at the time, specifically connectionist models (Smolensky, 1988; Hinton, 1990). System 1’s defining principle is that the statistical correlation between features in the experienced environment results in the strength of their associations in the mind. This system has other principles for Sloman, including that it is based on personal experience, that it processes automatically, and that it is fast and efficient. However, unlike for Epstein and many other dual-system theorists, consciousness is not a distinction between the two systems for Sloman, since system 1 is responsible for functions like fantasy and imagination.

Sloman stresses the similarity of system 1 to statistical learning, noting that just one outcome is not enough to form this system’s response to a stimulus (similar to Epstein’s experiential system). Associative strength can take into account all kinds of similarity between objects including superficial (dad and bad), semantic (dad and father) and temporal (dad and home likely occur in one’s environment at similar times). However, this sort of model on its own cannot account for immediate learning, as acknowledged by
connectionist researchers (McClelland, McNaughton & O’Reilly, 1995). These researchers proposed that immediate learning happens in the hippocampus. Without making such anatomical claims, Sloman employs system 2 to solve the same problem conceptually. This system takes a set of propositions presumed to be true and uses symbolic logic to derive novel inferences, harking back to much earlier attempts at artificial intelligence (Newell, Shaw & Simon, 1958). Specifically, the kinds of rules to which Sloman refers are causal rules (they always work) that can be, but are not necessarily, fully abstracted from any contextual cues.¹

Unlike for Epstein, there appears to be no interaction between the two systems in Sloman’s (1996) original proposal other than in competition for the response (system 1 does not influence subsequent system 2 processing and system 2 does not inhibit system 1). In a recent update to his theory, Sloman (2014) suggested that system 1 is likely heavily affective. He also renounced his theory’s division of the two systems, instead suggesting a mechanism for system 2’s inhibition of system 1, specifically, a leaky

¹ A distinction could be made between following rules and conforming to them. If a system follows rules, the rules must exist materially, and if they were changed, the system would behave differently. If a system simply conforms to rules, the rules do not necessarily exist anywhere, and the behavior of the system cannot be changed by changing the rules. For example, temperature in an apartment may fluctuate like a sinusoidal curve over time. This would be rule-following if there is a thermostat that enforces this to be true (if the thermostat’s settings were changed, the temperature would no longer fluctuate sinusoidally. It would be rule-conforming if it simply happens to be true due to the behavior of the external climate thus far (colder during the night, warmer during the day). Sloman’s system 2 is rule-following, not rule-conforming.
inhibitory mechanism. In his conceptualization, the more coherent and sensible system 2’s response, the less weight system 1’s response gets, and conversely, the more affectively potent and relevant system 1’s response is, the less weight system 2’s response gets. Although this appears to be an interaction, this sort of mechanism can still be seen as simple competition for the response, rather than a modification in processing. In fact, leaky competing accumulators, which are essentially identical to Sloman’s leaky inhibitory accumulators, are often seen as a plausible alternative to diffusion processes (Teodorescu & Usher, 2013). As a result, Figure 1 is a fair, though incomplete, representation of Sloman’s theory (because the figure excludes the inhibitory accumulators).

Sloman puts great emphasis on holding simultaneous contradictory beliefs as empirical evidence because he believes it indicates two distinct and competing sources of mental information that are likely systems. However, holding simultaneous contradictory beliefs is equally relevant for Epstein’s (1994) theory because both are parallel-competitive models. System 1 finds associations and suggests responses; system 2 simultaneously executes rules and sometimes suggests different responses. They compete for the overt response in the mind of the decision maker, as evidenced by the simultaneously held contradictory belief. The associative beliefs have the added advantage of being automatic, that is, persisting in the face of the decision maker’s attempts to suppress them. Simultaneously held contradictory beliefs appear to be the only evidence that is specific to parallel-competitive theories (as opposed to other dual-system theories).
Evidence that Epstein describes, such as correlations with individual difference measures of propensity for cognitive and experiential thought, and the existence of certain biases, does not directly support parallel-competitive (rather than other) dual-system models because such an individual difference measure is equally consistent with any other dual-system design that includes a more cognitive system 2 and a more experiential system 1 (for example, the default-interventionist models described below).

To the best of my knowledge, Mukherjee (2010) offers the only mathematical version of a parallel-competitive dual-system model. This model is based on the work of Rottenstreich and Hsee (2001; Hsee & Rottenstreich, 2004) who suggested that under affect-rich conditions, the curvatures of the value and probability weighting functions of Prospect Theory become more pronounced. Mukherjee suggested that system 2 (he calls it the Deliberative System) is effectively an expected value calculator, whereas system 1 (he calls it the Affective System) ignores probabilities, and simply averages outcome utilities (exponentiated dollar outcomes). The final value of a risky outcome for an individual is the weighted sum of system 1’s value and system 2’s value. The mathematical framework provided by this model, however, is content free, so the word “affective” can be replaced with “associative,” and this model would approximate Sloman’s (1996) theory (but not his new and less popular opinion that the systems interact; Sloman, 2014). It would be less consistent with Epstein’s model because that model allows system 1 values to influence system 2 values. This model is explained in
greater detail and tested in Chapter 3. It is an example of the simpler parallel-competitive model outlined in Figure 1.

**Default-Interventionist Models**

Default-interventionist models are often called dual-system models. In these models, system 1 works very quickly and suggests intuitive responses. Most of the time, intuitions provide ecologically valid responses that guide behavior, but in some cases, these responses are inappropriate, and system 2 may override the intuition and replace it with a more reflective response (see Figure 2). Usually, once the correct alternative is found by system 2, overriding system 1 is not problematic; according to default-interventionist theorists, the issue is activating system 2 in the first place. Models differ regarding the defining characteristics of either system, but most theorists agree on a multitude of characteristics that are correlated with system 1 and system 2 processing. Traditionally, many theorists endorsed the view that the systems are encapsulated. That is, if a process, such as logical thought about a statement’s validity, belongs to a system, then it must have all that system’s qualities, such as logical, analytic, slow and conscious (Epstein, 1994; Sloman, 1996; Kahneman, 2003; Stanovich, 1999). More recent formulations simply postulate that the properties are correlated (so a process in a given system will usually, but not always, have all of its corresponding properties; Evans & Stanovich, 2013).
Kahneman (2003; 2011) and Kahneman and Frederick (2002) provide a domain-general and mostly default-interventionist model based on the dual-system theories described above (e.g., Epstein, 1994; Sloman, 1996) and others (e.g., Gilbert 1989, 1991; Chaiken and Trope, 1999) that they apply to the domain of decision making. In this model, system 1 (called System 1) possesses several qualities; it is fast, parallel, automatic, effortless, associative and slow-learning (it requires many trials to instantiate a behavioral proclivity). System 2, (also called System 2 by these theorists) is described as slow, serial, controlled, effortful, rule-governed and flexible. In a marked difference from many other theories (including all the parallel-competitive and other default-interventionist models described here), both systems are evoked by language and both can process simulated rather than experienced circumstances. This theory’s systems are encapsulated (as described above). Although the papers that describe the model include plentiful descriptions by example, some of the mechanics are unclear. For example, the basic premise of Kahneman and Frederick’s model is that an intuitive response to a stimulus is
introduced quickly and if any subsequent processing occurs, then system 2 overrides and corrects the initial response (“System 1 quickly proposes intuitive answers to judgment problems as they arise, and System 2 monitors the quality of these proposals, which it may endorse, correct or override;” Kahneman & Frederick, 2002, pp. 3). However, the authors also mention that the two systems work simultaneously and compete for control of the overt response (“we assume that System 1 and System 2 can be active concurrently, that automatic and controlled cognitive operations compete for the control of overt responses;” Kahneman & Frederick, 2002, pp. 3). They do not reconcile this flexible dual activity with the earlier basic conceptualization that appears to fit system 1 and system 2 processes into a rigid, sequential timeline.

The basic structure of Kahneman’s model is consistent with the simplified model in Figure 2, but suggestions that the two systems compete are inconsistent with Figure 2’s model and would require additional arrows, including one from system 2 to the decision. Because in Kahneman’s view, system 2 comes up with its own proposals, rather than modifying system 1 proposals, it may also be appropriate to make the arrow originating from system 2 to the arrow between the intuition and the decision, rather than the intuition itself.

Stanovich (1999, 2005, 2009) does provide a detailed description of a default-interventionist dual-system model by adding several features to the generic description (shown in Figure 2). He used various terminologies over the years, including System 1 and 2, type 1 and 2 processes, the Autonomous Set of Systems (explained below) and the
Analytic Processing System. Three main similarities and differences exist between these sets of terminologies. First, he always stresses evolutionary roots. System 1 is evolutionarily old and maximizes behavior based on the goals of sub-individual units (i.e., genes). System 2 maximizes the goals of the individual.

Second, in more recent formulations, system 1 is no longer one system according to Stanovich (2009); it is now The Autonomous Set of Systems. Specifically, system 1 includes many sub-modules that can operate independently and are each responsible for different tasks (e.g., depth perception, likelihood judgment). Stanovich chooses to group these sub-modules because they all share some properties, like independence with respect to working memory and lack of cognitive decoupling (explained below).

Third, recent formulations have also suggested splitting system 2 into two components. In these formulations, one component can decide to override system 1 (the reflective mind) and the other component, the algorithmic mind, can implement the override chosen by the reflective mind (Stanovich, 2009). Although both components require working memory resources, Stanovich assumes that only the algorithmic mind is linked to general intelligence. The reflective mind, on the other hand, is linked to personal preferences and dispositions. Allowing an additional component that is involved in bias correction, but is not related to intelligence, allows Stanovich to explain why his work indicates that intelligence is uncorrelated with some biases (Stanovich & West, 2008). In addition, it explains why thinking dispositions have been linked to heuristics and biases above and beyond intelligence tests (e.g., Bruine de Bruin et al., 2007). In some ways, Stanovich’s
(2009) proposal is no longer a dual-system theory in that both systems are split into sub-components with different features. However, he calls it a dual-system theory and provides definitional properties that separate the two systems.

Two concepts are central to Stanovich’s dual-system view. The first is cognitive decoupling or hypothetical thought (Leslie, 1987), the ability to create not only representations of states of the world, but copies of these representations that can be manipulated in the mind. These decoupled representations are thought to be costly to maintain in terms of mental resources, but because they can be manipulated, they allow for mental simulations.

To best understand these concepts, imagine participants are solving many syllogisms (determining whether a set of logical links between statements is valid). For example, participants are asked whether the third statement below logically follows from the first two statements: “All fruits can be eaten. Oranges can be eaten. Oranges are fruits.” In fact, it does not follow; the syllogism is invalid, though the last statement is true. But participants often claim that invalid syllogisms with true conclusions are valid (this is called the belief bias; Oakhill, Johnson-Laird, & Garnham, 1989). Imagine you are trying to test a dual-system explanation of this effect, postulating that one system is responsible for determining validity whereas the other is responsible for determining truth of the conclusion.

Although the conclusion that oranges are fruits is true, to surmise the validity of the syllogism, it is useful to consider the possibility that oranges are not fruits. Finding that
this alternative conclusion is also consistent with the premises may lead the reasoner to correctly conclude that the premises do not justify the conclusion. Stanovich’s view is that cognitive decoupling is one of the primary functions of system 2 and it happens in the algorithmic mind in particular (and not the reflective mind) because cognitive decoupling is a function of ability, not preference toward reflection. He holds that most intelligence measures, executive functioning measures, and working memory measures test this ability.

The algorithmic mind is also capable of non-hypothetical thought (i.e., thought that does not require cognitive decoupling). Specifically, it is capable of what Stanovich calls Serial Associative Cognition. This captures any deliberative thought directly tied to a percept or a fact about the world (rather than a counterfactual). In the case of the syllogism, this might be conscious, verbalized reasoning explaining why and how the premises do not falsify the conclusion. For example, statements like “All oranges can be eaten, and all fruits can be eaten, so the conclusion is valid” do not add any representations that are not already in the syllogism, and yet they require cognitive effort to verbalize. According to Stanovich, the algorithmic mind, although enough to solve an abstract syllogism (e.g., if all A’s are B’s and Some B’s are C’s …), is insufficient for syllogisms that trigger the belief bias. A syllogism may be solved incorrectly due to a failure in a calculation of the algorithmic mind, or it can be incorrectly solved due to belief bias (or any other bias); this latter error is an error of the reflective mind, not the algorithmic mind.
The second concept specific to Stanovich’s view is the idea of mindware. He suggests that none of the components of either system exist without content. The various components of system 1 (the autonomous set of systems) exist either through evolutionarily conveyed knowledge or through “over-learned” “or “tightly compiled learned” information like associations or motor skills. This means that, although every person possesses system 1, a specific person might not have an over-learned module for a specific task. Therefore, system 1 may suggest different responses to different people in the same situation. For example, to a person who has never heard of an orange before and does not know it is a fruit, system 1 will not suggest a response consistent with belief bias. But, because most of us cannot help but think “fruit” when we are told about an orange (even though it is irrelevant to the validity of the syllogism; Stanovich would say we over-learned this association), this belief bias influences responses. Because the reflective mind of system 2 is responsible for controlling system 1, if the belief bias influences the response, it is the reflective mind’s error. Similarly, the algorithmic mind (also part of system 2) acts on learned strategies and production systems. For example, the validity of a syllogism may be much easier to determine for someone who has taken a formal logic class and learned tools to help solve such problems. Finally, the reflective mind takes into account prior beliefs, goals, and general knowledge. For example, a participant who has been to a previous psychological experiment in which s/he was tricked by an experimenter may address the syllogism more carefully and be more likely to think about it in detail and solve it correctly. Although Stanovich’s original theory is
largely consistent with Figure 2, the elaborations on this theory, including the autonomous set of systems and the difference between the algorithmic and reflective minds, are not reflected in that simplified representation. However, there need not be an arrow from system 2 to the decision for any version of Stanovich’s model, because system 2 only manipulates system 1 proposals, but does not make its own proposals.

In another default-interventionist theory, Evans has proposed and advanced what he calls the heuristic-analytic, dual-system theory (Evans, 1984, 1989, 2006), in which the heuristic system is system 1 and the analytic system is system 2. Like in Stanovich’s approach, hypothetical thought is central in Evans’ approach. The novelty of this approach is that hypothetical thought is conducted by a loop of system 1 proposals of hypothetical representations and system 2 evaluations of these proposals, instead of one heuristic response that is either overridden or not.

At each step of the loop, system 1 comes up with a most likely representation. If system 2 is satisfied, the loop ends, and this representation is used for judgment and decision making (this often happens on the first iteration if motivation is low). But if system 2 is not satisfied, system 1 continues generating representations until system 2 finds one acceptable. So in this model, even if system 2 overrides system 1’s initial representation, the representation that ultimately leads to a decision is still generated by system 1.

System 2 encodes and tests “mental models,” which for Evans are conditional statements coupled with subjective probabilities (e.g., I am 90% confident that if I work, I will be paid). Such conditional statements are the basis of hypothetical thought, which is
essential for system 2. Evans, Over and Handley (2003) proposed three principles that they believe govern hypothetical thought. First, their singularity principle suggests that only one proposed representation is considered by system 2 at a time. The relevance principle suggests that system 1 suggests the most relevant (plausible and/or probable) representation in the current context. Finally, the satisficing principle echoes Simon’s (1972) idea that people are cognitive misers who employ a bounded rationality: System 2 will reject a proposal only if it is unacceptable (not merely if it is suboptimal).

System 2 has several features in Evans’ interpretation: It is slow, sequential, controlled, requires working memory resources, has inhibitive properties, and is responsive to verbal instructions. System 1 has the opposite features on all these dimensions. However, the defining feature of system 2 is that it requires working memory resources (and the defining feature of system 1 is that it does not; Evans & Stanovich, 2013). Although these features are similar to those offered by Stanovich and by Kahneman, the architecture of this model, specifically the loop, is distinct. In this treatment, all proposals, even those that are considered deliberative, are generated by system 1, hence Figure 2 is most representative of this model and it does not appear that any changes are required to Figure 2 to account for this model, whereas other models do not explicitly exclude the ability of system 2 to influence decisions through avenues other than inhibition of intuitions.

One of the primary means used to test (and often to support) all default-interventionist models is the Cognitive Reflection Test (CRT; Frederick, 2005). The CRT is intended to
measure the ability to reflect on a default response, and then to inhibit it, so it is thought to be the theoretically most appropriate individual-difference measure for differentiating default-interventionist theories from other dual-system theories (because inhibition of intuition is a key aspect shared by default-interventionist theories and that is distinct from other dual-system theories). Higher scores on this test have been shown to correlate with less bias on various decision-making tasks, providing evidence for the validity of the default-interventionist perspective in explaining judgment and decision making biases (Frederick, 2005; Toplak, West & Stanovich, 2011; Toplak, West & Stanovich, 2014).

Chapter 2 is intended to more critically evaluate whether correlations with the CRT indeed constitute evidence for default-interventionist models.

**Interactive Models**

Unlike the models discussed so far, some models of decision making are fully interactive. These models are generalizations of parallel-competitive and default-interventionist models that emphasize the complex interactions between the two systems. Unlike in Epstein’s (1996) theory, for example, multiple ways the systems interact are described in more detail, with specific interactions between the systems outlined, though other, yet unspecified interactions are usually also allowed. An emotion or association can spark a thought that sparks another emotion, etc. These interactions are more general than a leaky inhibition process, because in addition to inhibiting each other’s response, the systems can influence each other’s processing before competition for a response begins. For example, when making a decision about whether to take a risky gamble, a negative
emotion about the gamble (“a $50 loss is really bad”) can trigger a negative thought (“I only have $50 in my wallet”), but it can also trigger a positive thought (“but the win is ten times larger”), reflecting that emotions of either valence can lead to thoughts of either valence. When deciding whether to smoke, the presence of a graphic image on a warning label can elicit an emotional reaction which subsequently causes the smoker to think harder and learn more about presented health risks (Evans, et al., 2015).

Examples of these models include Loewenstein et al.’s (2001) and Slovic et al.’s (2004) models. These researchers see two routes to risk perceptions and decisions: through feelings and through analysis. Risk as analysis is similar to previous models, like Prospect Theory, which suppose the risk assessor undertakes some sort of subjective calculus, similar to that of system 2 in Sloman’s model. Risk as feelings, on the other hand, includes not only feelings associated with the stimulus (as might be expected in Sloman’s and especially in Epstein’s system 1), but also includes incidental emotional states (most other theories do not assign a significant role to incidental emotions). The two routes, independently shaped by stimuli, and by each other, work interactively to guide behavior. These authors provide plenty of evidence that even bounded rationality does not account for human decision making including risk perceptions; emotions appear to play a central role. For example, Damasio (1994) showed that emotion-based somatic markers are integral to learning effective behavior in gambles.

Damasio’s (1994) own model can also be classified as interactive. In Damasio’s view, emotion and reason are entwined rather than separate. Human experiences first are
encoded in an emotional way, then, given a long and intense enough exposure, an experience may become consciously recognized. Conscious recognition may modify the experienced emotions, but not to a great degree. Instead, it works with the emotional information in order to draw effective conclusions. Damasio draws on evidence from neuroscience, for example a patient with damage to the ventromedial prefrontal cortex, to argue that some brain areas are responsible for the role of emotion in decision making, but not performance on cognitive tasks. In Damasio’s view, emotions and cognitions are separable, but they are inextricably linked in effective decision making. Damasio, unlike many dual-system researchers, argues that in most cases, even higher order consciousness is largely emotional. In other words, even parts of the brain responsible for purely cognitive tasks do not operate separately from the emotional brain areas in most situations.

Although these interactive models may accurately categorize the complexity of many real, important decisions, it is difficult to mathematically fit these models to data without on-line measurements of the two systems, like brain scans or skin conductance measures. If they can interact in unspecified ways, any manipulation that has an expected effect on one system may have an unexpected effect on the other due to one of the interactions. As a result, any pattern of the resulting behavioral data can be explained by some form of interaction between the two systems. However, experimental evidence can support various interactions between the two systems piece-by-piece. For example, experimental studies support the idea that incidental affective states can influence the type of reasoning
employed. Positive affect, for example, can induce more creative and open-ended problem solving (Isen et al., 1985; Isen, Daubman & Nowicki, 1987) and can influence the subjective likelihoods of outcomes (Johnson & Tversky, 1983).

**Importance of working memory in dual-system models**

Each of the models described above and most other dual-system models are tied to working memory, whether explicitly or implicitly. Some theorists, like Evans and Stanovich (2013), explicitly mention that working memory is an important, and possibly unique, mechanism to enable an empirical test of their theories. An individual’s working memory capacity is usually measured or manipulated through time pressure, or a concurrent task. However, working memory as an empirical test is implicit in many other theories. Kahneman and Frederick (2002) for example, subscribe to the idea that system 2 is controlled and serial, suggesting that it slowly processes one stimulus at a time. Epstein (1994) and Sloman (1996), on the other hand, implicate language in system 2. Since the phonological loop, a subsystem of working memory, is thought to be necessary for language (Baddeley, Gathercole, & Papagno, 1998), a manipulation that would tax working memory should make system 2 processing less likely in their models as well. Given a heavy enough manipulation, language-based cognition should be inhibited even if the manipulation does not tax the phonological loop specifically because the central executive is required for all working memory processes. Studies exist employing competing working memory tasks or time pressure conditions while having participants
complete standard reasoning bias tasks. These studies indeed show that there is greater bias with greater working-memory load, whether due to a concurrent task or time pressure (De Neys, 2006a, 2006b; Roberts & Newton, 2001). Chapter 3 builds on the tradition of using working-memory load to examine dual-system models’ predictions in risky choice.

Alternatives to and criticisms of dual-system models
Recently, several prominent criticisms of dual-system theories have emerged (Keren & Schul, 2009; Gigerenzer, 2010; Kruglanski & Gigerenzer, 2011; Frank, Cohen, & Sanfey, 2009). The criticisms can be placed in five categories (see Table 2). They will be addressed in the order in which they are presented in the table. The first two criticisms (namely that evidence for two systems is also consistent with one system and that more than two systems are necessary) come from proponents of single and multiple system models, respectively, and are addressed after brief reviews of these types of models. The next three (that dual-system models are inconsistent with each other, they do not make empirical predictions, and they are not clear or specific enough) are then considered.
Table 2. Common criticisms of dual-system theories

<table>
<thead>
<tr>
<th>Criticism</th>
<th>Source(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence for two systems is also consistent with one system</td>
<td>Osman, 2004; Kruglanski, 2014; Kruglanski &amp; Gigerenzer, 2011</td>
</tr>
<tr>
<td>More than two systems are necessary</td>
<td>Squire, 2004; Frank et al., 2009</td>
</tr>
<tr>
<td>Dual-system models are inconsistent with each other</td>
<td>Keren &amp; Schul, 2009</td>
</tr>
<tr>
<td>Dual-system models do not make empirical predictions</td>
<td>Keren &amp; Schul, 2009; Kruglanski &amp; Gigerenzer, 2011</td>
</tr>
<tr>
<td>Dual-system models are not clear/specific enough</td>
<td>Frank et al., 2009; Keren &amp; Schul, 2009; Gigerenzer, 2010</td>
</tr>
</tbody>
</table>

**Review of single continuum models**

In addition to dual-system models, there exist single continuum models that are constructed quite similarly and make similar predictions about thoughts and behaviors (e.g., Gilbert, 1999). For example, a single continuum model might suggest that when thoughts are more language-based, they will also be less automatic because automaticity and need for language might lie opposite ends of the continuum. Furthermore, in a default-interventionist model, system 2’s intervention does not have to be all-or-nothing; a response can be guided by the intuition to some degree even if an inhibition is taking place...
place. In a parallel-competitive model, neither of the systems has to “win out” and take control of the response. Both systems can have an influence. The difference between dual-system and single system accounts, then is that the dual-system models assume a mechanism such as competition or intervention by which this range of behavior comes about, whereas a single continuum model makes no such assumptions, but simply states the correlated characteristics, examples of which are described below. As a result, evidence for dual systems can be consistent with a single continuum model (e.g., see below for how Kruglanski’s (2014) Unimodel might account for evidence used to support dual systems).

**Cognitive Continuum Theory**

Hammond (1981) proposed a theory, called Cognitive Continuum Theory (CCT), intended to summarize much of the work-to-date in judgment and decision making. He made intuition and analysis two endpoints on a continuum of modes of processing. In CCT, the area between the ends is referred to by laypeople as common sense and referred to by Hammond as quasi-rationality. In CCT, cognitive tasks that most individuals undertake in most situations, including psychological experiments, are somewhere between the endpoints. Like most dual-system approaches, CCT maintains that the various properties are highly clustered. For example, intuition is under low cognitive control, is rapid and unconscious, whereas analysis is under high cognitive control, slow and conscious.
In Hammond’s theory, the analytic end of the spectrum corresponds to normative solutions to complex problems. Depending on the complexity, such a solution may involve calculus, statistics, simulation, and information from scientific studies. Processing on this end concentrates on finding the true solution to the problem. Hammond believes that, although other approaches have correctly identified analysis, intuition has been poorly understood. The intuitive end of the spectrum in CCT substitutes cues that are readily available in the environment and are correlated with the answer for any sort of formal calculation or deliberation. The substitution is very similar to perceptual substitutions and therefore can be analyzed using a range of methods proposed by Brunswickian psychology (Hammond & Stewart, 2001). For example, a Brunswickian analysis would show that the size of a familiar object on the fovea can be substituted for distance in most circumstances. In the same way, this analysis can show that a tennis player with better name recognition is likely to win a match (in support of the Recognition Heuristic; Goldstein & Gigerenzer, 2002). Similarly, Kahneman (2003) built the rationale for his dual-system model based on a very similar analogy to perception. CCT is much more detailed in terms of its mechanics than most dual-system approaches (e.g., Epstein, 1994; Kahneman, 2003). It names a number of conditions that would shift a reasoner on the continuum. As a result, it is able to make many specific empirical claims that dual-system approaches do not currently make. For example, in addition to the usual predictions that reducing available processing resources should move
processing away from analysis to intuition, CCT suggests other experimental conditions that can induce analysis: presenting stimuli as dichotomous (rather than continuous, because dichotomous options are easier to think about carefully), in numbers (rather than pictures), with precise (rather than approximate) values. In addition, it makes predictions about response types that characterize analytic processing. For example, intuition is characterized by normal error distributions, and analysis is characterized by irregular, highly peaked, error distributions. However, because the predictions are also entirely consistent with dual-system models (albeit more detailed), it is difficult to test against or compare with dual-system models. It is not clear why dual-system models remain much more popular than CCT. It is possible that it is due to their longer history, which according to some goes all the way back to Aristotle (Keren & Schul, 2009), or due to the fact that two systems are more intuitive to researchers than a single mode.

*The Unimodel*

Another alternative to dual-system approaches is the Unimodel proposed by Kruglanski (e.g., Kruglanski et al., 2003). The Unimodel proposes alternative explanations to phenomena traditionally ascribed to dual-system models in a variety of domains. The version most relevant to the present thesis is that proposed by Kruglanski and Gigerenzer (2011) regarding decision making. The core principle of this theory is that cognition is based on a multitude of rules, with association simply being one of the rules (e.g., “choose the option with the greater association to the stimulus”). These rules vary in terms of their evolutionary age, processing ease, speed of application and many other
characteristics. Rule choice is governed by individual memory and perceived ecological relevance (e.g., of the rules that are readily available in memory, the most ecologically valid are chosen). When the perceived ecological validity of several rules is similar, the decision of which to apply may be difficult. When a rule is difficult to apply (for example, because it has not been routinized), individual differences in motivation and cognitive capacity will determine the success of its application.

The Unimodel can account for much of the evidence that has been used in support of dual-system models. For example, the belief bias in syllogistic reasoning example about oranges being fruits outlined earlier in the text can be explained as the application of two rules. The first rule is that accessible prior beliefs are correct, implying that oranges are fruits is valid. The second rule that involves the determination of the validity of the syllogism is more difficult to apply, and therefore less likely to be used and especially under time pressure or cognitive load.

Compared to CTT, the Unimodel’s predictions are less clear because it is impossible to enumerate all rules. Like CCT, the Unimodel can account for a range of phenomena associated with dual-system views, such as simultaneously held contradictory beliefs and differences in performance under time pressure and cognitive load. Also like CCT, the Unimodel is difficult to compare to dual-system models because the predictions it makes are so similar.

The contention that evidence used in support of dual-system theories is consistent with single-system theories was also raised by Osman (2004). Unlike most dual-system critics
Osman reviewed several dual-system theories, one at a time, and explained why the evidence each presented is also consistent with a single continuum. His arguments were similar to those of Hammond (1980), suggesting that a range of deliberativeness can account for the same findings. In the context of memory, Ratcliff, Van Zandt, and McKoon (1995) made a similar argument quantitatively. At that time, a method developed by Jacoby (1991) for measuring two processes’ contributions to retrieval was popular. Given retrieval data, the method was able to assess a conscious recollection process and an unconscious familiarity process. Ratcliff and colleagues (1995) generated memory data using a dual-system model and a single-system model and found that Jacoby’s (1991) method implied that two processes contributed even when data were generated from a single process. This argument suggests that quantitative dual-system theories can be vulnerable to the same criticisms as verbal theories: just because a dual-system model fits the data does not mean the data are any less consistent with a single-system model.

Although all these authors’ arguments suggesting that dual-system evidence is often consistent with a single-system explanation are valid, they do not constitute evidence that single-system theories are correct, or even preferable. One way to assess which model is superior when both fit the evidence equally well is parsimony. But at least for verbal models, it is unclear whether single-system theories are more or less complex than dual-system theories. As a result, parsimony cannot be used to select the better model without formalizing both and assessing their complexity, for example, by counting parameters.
In sum, the single-system models can be seen as viable alternatives to dual-system models. However, because both frameworks are so general in decision making research, each can likely account for similar findings. Oftentimes, the same empirical evidence can be interpreted as supporting single-system models (like CCT or the Unimodel) and dual-system models at the same time, since both contain a prominent role for working memory and inhibition of intuitions. The rest of the proposal uses the language of dual-system models because these models have been more widely used in the recent decision literature.

**Review of multiple systems models in relation to dual-system models**

There exists a significant amount of evidence supporting multiple systems in memory. This research has direct implications for judgment and decision making because judgments and decisions often involve processing of memories (e.g., memories of previous experiences and personal values). Furthermore, current evidence suggests that online processing (such as what has been referred to as both system 1 and 2 processing) can be seen as activated long term memory (e.g., Cowan 2005; Postle, 2006; Ruchkin et al., 2003). As a result, if credible evidence of multiple systems in long term memory exists, it may imply the existence of multiple systems in judgment and decision making. In this section, evidence for multiple systems in memory is reviewed, and then it is related to theories of decision making.
Evidence for multiple systems in memory

A host of research points to separable systems of memory. The history of the exploration of multiple systems of memory is described by Nadel (1994), Squire (2004), and others. Early evidence of dissociations came from amnesic patients with lesions in the hippocampus. They were shown to be capable of procedural memory (memory of how to do something), but completely incapable of declarative memory (memory that something occurred). In one elegant demonstration, Cohen and Squire (1980) showed that amnesic patients can learn the skill of mirror reading, without retaining memory for the actual words they read, or even the sessions during which they learned the skill. Initially, it was thought that such evidence could be interpreted to mean that memory for procedural and motor tasks is not really memory, or at least is not memory in the same sense as declarative memory is. After all, knowing how to ride a bicycle or write is generally considered ability, not a memory. Other research showed that some cognitive learning abilities more traditionally associated with memory also remain intact in amnesic patients (like category learning or grammar learning; Knowlton, Ramus & Squire, 1992; Knowlton & Squire, 1996). Later research showed that amnesic patients also exhibit perceptual (Hamman & Squire, 1997) and conceptual priming (Levy, Stark & Squire, 2004), indicating that priming, which is thought to be a memory process (Tulving & Schacter, 1990), involved a type of memory different from declarative memory. This research indicated that aspects of recognition memory were intact in these patients and that they were not missing all types of memory, but rather just one type.
Initially, researchers stopped short of proposing multiple memory systems, since differences between normal and damaged-hippocampus patients could be explained by other factors, such as a deficit in retrieval. Specifically, it is possible that a single memory system existed, in which the hippocampus was responsible for retrieving stored memories (Warrington & Weiskrantz, 1970), and therefore hippocampus lesions simply inhibited retrieval, though the memories existed. Eventually, others began to propose memory dichotomies. The first such suggestions came from behavioral neuroscientists working with rats who suggested that fast single-trial hippocampal learning about environments (in the case of rats in mazes) is different from slow learning typical in traditional behavioral paradigms (e.g., Nadel & O’Keefe, 1974). Soon, dichotomies like declarative vs. procedural memory (Cohen & Squire, 1980), implicit vs. explicit memory (e.g., Graf & Schacter, 1985), and complex and simple associations (Sutherland & Rudy, 1989) among others, became popular. Although these authors did not agree on the specific of the characteristics of the systems, they did agree that at least two systems were a good explanation for findings in memory research. By the 1990’s, most of the field shared the perspective that more than one system was necessary, although some disagreed (e.g., Roediger, Weldon & Challis, 1989; Masson & Macleod, 1992) and no consensus existed on what a system was.

However, the dichotomies based on the hippocampus soon proved not flexible enough. Affective learning, such as that which happens in fear conditioning, depends on the amygdala (e.g., LeDoux, 1993), whereas habitual and procedural learning take place in
the striatum and neostriatum (Mishkin et al., 1984). Consequently, some researchers proposed multiple memory systems. Squire (2004) held on to the separation between declarative (which includes episodic) and non-declarative memory, but changed ‘non-declarative’ to a label for a set of systems composed of procedural, perceptual, conditioning, and non-associative learning; he did not consider it a unitary system. This classification is similar to Stanovich’s (2009) rebranding of his system 1 as The Autonomous Set of Systems in that it groups non-declarative systems together similar to how Stanovich grouped the autonomous systems. The quest for identifying separate brain systems appears to have lost popularity since that time, with many researchers concentrating on the functions of very specific brain regions (Koban & Pourtois, 2014).

**Possible implications of multiple memory systems for decision making**

Some researchers, however, are interested in how multiple brain systems might relate to decision making. Frank, Cohen, and Sanfey (2009), for example, outlined a multiple-systems view of decision making based on multiple systems in the brain. The brain regions they describe correspond to memory systems identified in the research described above both by location (e.g., amygdala, hippocampus) and function (one-shot vs. experiential learning), although Frank and colleagues concentrate on what happens to decision making when these systems are activated, instead of any effects on memory. This account is formulated as an explanation of how traditionally accepted contrasts (emotional vs. cognitive, automatic vs. controlled, and episodic vs. associative) can be mapped onto a multiple systems view. Cognitive and controlled processing (e.g., the use
of system 2 in decision making) appears to be handled by the prefrontal cortex, whereas associative and automatic processing (generally considered system 1) are facilitated by the basal ganglia. Emotional processing (associated with system 1 by most) occurs in the amygdala. The hippocampus handles episodic memory (associated with system 2 because it is required for verbal recollection). They also describe the mechanics of the interactions between some of these systems.

Similar to Stanovich’s proposal of rebranding system 1 as The Autonomous Set of Systems and similar to other multiple system proposals (e.g., Schacter & Tulving, 1994), these researchers propose that emotional, associative, and automatic processing are not related and do not constitute one “non-deliberative” system. According to Frank et al., the prefrontal cortex, which corresponds to the functions of system 2 in many theories because it is related to working memory function, not only inhibits processing (as suggested by default-interventionist models), but it also amplifies relevant signals in other areas. It is less similar to Epstein’s model, in which the primary mechanic is system 1 influencing system 2. The mathematical models proposed by Frank and others (Daw, Niv & Dayan, 2005; Frank, Moustafa, Haughey, Curran & Hutchinson, 2007; Miller & Cohen, 2001) rely on neuroscientific data in order to be fit.

Although the combination of Frank, Cohen, and Sanfey (2009) and the evidence summarized in the previous section suggest that two systems are insufficient for describing human brain function, it is less clear how their evidence relates to describing behavioral decision making. Many contrasts they described could be reframed as the
prefrontal cortex (aligned with system 2) modulating processing of other brain areas (aligned with system 1); this description is quite similar to a default-interventionist proposal. Thus, in terms of judgments and decisions, although the distinctions between the other (non-prefrontal cortex) brain areas may be important, contrasts that are consistent with dual-system theories may also remain valuable.

In addition, the inability to account for the comprehensive range of human cognition is not a weakness specific to dual-system models. Currently, any multi-system model also needs to include a poorly understood “placeholder” system to acknowledge that it is not a complete description of human cognition or brain function because the search for systems has by no means been declared complete. In Squire’s (2004) memory model, for example, this “placeholder” system is the non-associative learning system, which is not specifically described. The role of placeholder is sometimes relayed to the procedural or perceptual or “some other, as yet little understood, memory system” (Tulving, Schacter & Stark, 1982, p. 341). Since decision making involves cognition and memory, the same criticism applies to dual-system models of decision making. Some decisions (e.g., non-deliberative “decisions” made out of habit) will always need to be delegated to similar procedural, perceptual or other poorly understood decision systems. In a single continuum model, the same problem can be handled with yet undiscovered rules.
Other criticisms of dual-system models

Single-continuum and multiple-system models provide some arguments against dual-system models, but other criticisms exist. This section discusses three additional criticisms of dual-system models (see Table 2).

First, critics often argue that dual-system theorists do not share the same global dual-system model that includes the same properties in all models (Karen & Schul, 2009). Thus, the theorists are inconsistent with each other. Second, they argue that this generic model (an amalgamation of dual-system accounts) makes few predictions about behavior (Keren & Schul, 2009; Kruglanski & Gigerenzer, 2011). It is true that some dual-system models contradict each other. For example, default-interventionist models argue that system 2 does not begin operating until system 1 is finished, but parallel-competitive models claim processing is simultaneous in the two systems (Evans & Stanovich, 2009). However, dual-system theorists also do not usually claim to agree with all other dual-system theorists. And, many theorists never subscribed to the generic model (Evans & Stanovich, 2013).

Most theorists’ more nuanced models do make specific predictions (e.g., regarding the CRT and working memory), that may or may not be supported by data. For example, some models make predictions regarding how individual differences should correlate with decision-making biases that have been tested (e.g., Frederick, 2005; Toplak, West & Stanovich, 2011). Chapter 2 is intended to critically assess the sort of evidence that one widely-used individual difference measure, the CRT, provides for dual-system theories.
Other models make predictions regarding experimental manipulations, such as working-memory load. These are tested less frequently in decision making, but several studies exist (Hinson, Jameson, & Whitney, 2003; Whitney, Rinehart, & Hinson, 2008; Figner et al., 2009). For example, Whitney and colleagues tested but failed to find support for one decision bias increasing with a greater working-memory load (Whitney et al., 2008). Specifically, they showed that, although likelihood to gamble generally decreases under a concurrent working-memory load, it decreases equally regardless of whether a gamble is framed as a loss or gain; in other words, working-memory load did not moderate framing effects. Although other studies have found support for dual-system predictions for other biases like syllogistic reasoning (e.g., DeNeys, 2006b), this study found no effect of working-memory load on a framing bias.

A final contention is that dual-system models are often not specific or clear enough. In particular, dual-system models rarely, if ever, make quantitative predictions (Frank et al., 2009). It is true that dual-system models are rarely quantitative (this is also true of single-continuum models). In order to address this concern, Chapter 3 reviews one quantitative version of a parallel-competitive model and tests it.

**Summary**

This introduction summarized some of the research on dual-system theories used in reasoning and decision making with an aim of classifying them into discrete categories and reviewing alternative models and common criticisms. Some of the most substantial criticisms of dual-system models argue that the models are vague and do not make
specific predictions. The next two chapters are intended to approach these theories empirically and offer novel tools for researching them as well as to provide empirical evidence for or against two of the specific categories of models reviewed here. In Chapter 2, predictions from a default-interventionist model are drawn and tested whereas in Chapter 3, a parallel-competitive model is tested in the form of a mathematical model. The studies in Chapters 2 and 3 provide empirical tests of two of types of models discussed in this chapter and thereby attempt to address the criticisms also discussed here.

**Transition to Chapter 2**

Chapter 2 examines the CRT, a scale that is purported to measure individuals’ reflectiveness, and that has been shown to correlate with several decision-making biases. These correlations have been argued as evidence in support of dual-system theories, especially those of the default-interventionist variety (Frederick, 2005). As a result, examining the CRT closely is an important step to finding evidence for or against this type of dual-system theories. Since the CRT is also purported to be a test of the ability to inhibit biased intuitions (Toplak et al., 2011), it may be relevant to other dual-system theories, like Epstein’s (1994) CEST, in which biased intuitions are a hallmark of system 1.

Chapter 2 is a reprint of a published paper (Sinayev & Peters, 2015, *Frontiers in Psychology*) in which we replicated and reanalyzed correlations between the CRT and various decision-making biases. In particular, in two studies, we disaggregated CRT scores into the separate components of cognitive reflection and numeric ability using
proportions and a psychometric method called IRTrees. We then demonstrated that correlations with the CRT do not necessarily indicate that inhibiting intuitions per se is important to reducing decision-making biases (e.g., conjunction fallacies) and negative outcomes (e.g., taking predatory loans). Numeric ability, measured using the CRT or a numeracy scale, instead accounted for the CRT’s ability to predict more normative decisions (a subscale of decision-making competence, incentivized measures of impatient and risk-averse choice, and self-reported financial outcomes) even when controlling for many relevant confounding variables like income and intelligence. Although default-interventionist dual-system predictions suggest that as a measure of inhibition of intuition Cognitive Reflection should correlate with less biased decision making, Cognitive Reflection contributed no independent predictive power. Thus, Chapter 2 findings indicate that one common method that has been touted as empirical support for default-interventionist dual-system theories may not truly provide such support. The final paper follows as Chapter 2.
Scores on the three-item Cognitive Reflection Test (CRT) have been linked with dual-system theory and normative decision-making patterns (Frederick, 2005). In particular, the CRT is thought to measure monitoring of System 1 intuitions such that, if cognitive reflection is high enough, intuitive errors will be detected and the problem will be solved. However, CRT items also require numeric ability to be answered correctly. In two studies, we examined whether the Cognitive Reflection Test (CRT) was predictive of superior decision making because it measures the ability to check intuitions and/or the ability to solve numeric calculations.

The Cognitive Reflection Hypothesis

The CRT is a popular three-item test (Frederick, 2005) thought to assess cognitive reflection because the items bring to mind intuitive but wrong solutions that have to be overridden. The prototypical CRT problem is the bat and ball problem: “A bat and a ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?” The response “10 cents” is thought to come to mind for most, if not all, people, and many
people answer “10 cents.” Some people realize that the intuitive response is incorrect, however, and researchers have believed that calculating the correct answer is straightforward at that point: “catching [the] error is tantamount to solving the problem” (Frederick, p. 27). Kahneman (2011) called the bat and ball problem “a test of people’s tendency to answer questions with the first idea that comes to mind, without checking it” (p. 65). Consistent with this view, we define Cognitive Reflection as the tendency to check and detect intuitive errors, and call the hypothesis that it is the important aspect of the CRT, the Cognitive Reflection Hypothesis.

In support of the Cognitive Reflection Hypothesis, Frederick (2005) briefly noted several pieces of unpublished evidence. In particular, people who responded correctly sometimes wrote the intuitive answer in the margin and described thinking about the intuitive answer in verbal reports, indicating that the intuition did come to mind. People who answered incorrectly thought the bat and ball problem was easier than those who answered correctly (incorrect responders judged the proportion of others who answered correctly to be higher than correct responders did), indicating that those who responded intuitively were unaware that the intuition was wrong. Inconsistent with this reasoning, however, De Neys, Rossi, and Houdé (2013) found that correct responders were more confident about their responses than incorrect responders. Frederick (2005) also noted that some people who perform badly on the CRT nonetheless are able to solve similar problems that do not have incorrect intuitive solutions (e.g., “a banana and a bagel cost 37 cents. The banana costs 13 cents more than the bagel”). However, Bourgeois-Gironde and Van der Henst
(2009) subsequently demonstrated that most people answer these problems incorrectly anyway (58% incorrect; see also Mastrogiorgio & Petracca, 2014).

Alter, Oppenheimer, Epley, and Eyre (2007) provided evidence consistent with the CRT assessing an increased tendency to check intuitions. In particular, they found that participants who read the CRT in a degraded font (which presumably increased information processing) answered correctly more often than participants who read it in a normal font. However, the effect was not limited to tasks that require checking and inhibiting intuitive responses. Diemand-Yauman, Oppenheimer, and Vaughan (2011) demonstrated that disfluent fonts improved performance on a wide range of tasks (including ones with and without intuitive responses). Their results indicate that the improvements in CRT performance may have been due to some other process such as a more general increase in deliberation rather than a specific increase in intuition checking.

In sum, although some evidence exists that the CRT measures cognitive reflection, the same evidence is also consistent with it measuring other constructs.

**Dual-System Explanation**

To explain his findings, Frederick (2005) invoked a dual-system model of decision making. In it, intuitive System 1 processes are quick and effortless whereas deliberative System 2 processes are slow and controlled. System 1 quickly makes an intuitive response available in decision making; System 2 then may check the response and engage in further reasoning if an error is detected (Kahneman, 2003; Kahneman & Frederick,
Importantly, System 2 is activated only after System 1 processing is complete. This temporal rigidity distinguishes it from dual-system explanations of judgment and decision making that posit more interdependencies between the two systems (Slovic, Finucane, Peters, & MacGregor, 2004; Loewenstein, Weber, Hsee, & Welch, 2001). Many biases are said to occur due to System 1’s incorrect intuitions, so that people who check their intuitions (e.g., those scoring high on the CRT) should be less biased decision makers.

Consistent with this prediction, several studies have found correlations between the CRT and decision biases. In his original paper, Frederick (2005) found that people with lower CRT scores tended to be more impatient and risk averse, therefore failing to maximize expected utility. Oechssler, Roider, and Schmitz (2009) also found that people higher on the CRT were less likely to commit conjunction fallacies and conservatism in probability updating. Other researchers have found expected CRT correlations with probability updating, base rate neglect, and under/over confidence (Hoppe & Kusterer, 2011), regression to the mean, Bayesian reasoning errors, and framing effects (Toplak, West, & Stanovich, 2011), performance on Wason selection and denominator neglect tasks (Toplak, West, & Stanovich, 2014), and moral judgments (Paxton, Ungar & Greene, 2012; Royzman, Landy & Leeman, 2014). That CRT scores correlate with fewer judgment and decision biases has been interpreted as indicative of bias avoidance requiring one to check and correct intuitions and, therefore, as support for a dual-system explanation of decision making (Thaler & Sunstein, 2008; Kahneman, 2011).
Each of these researchers assumes that differences in CRT performance indicated differences in the ability to detect and correct incorrect intuitions (i.e., the Cognitive Reflection Hypothesis). They also implicitly assume that numeric ability is an irrelevant detail when it comes to solving CRT and related problems. Contrary to this view, however, Baron, Scott, Fincher, and Metz (2014) recently found that traditional CRT problems have no more predictive power with respect to moral preferences than similar arithmetic items without intuitive answers. These findings suggest that numeric ability may be important to CRT performance.

The Numeracy Hypothesis

Other researchers include CRT items in measures of numeric ability, implying that the CRT is not substantially different from other math tests (Weller et al., 2013). In fact, four of the five published studies employing exploratory or confirmatory factor analyses concluded that CRT and other numeracy items load on the same factor (Låg, Bauger, Lindberg, & Friborg, 2014; Baron, et al., 2014; Weller et al., 2013; Study 1 of Liberali, Reyna, Furlan, Stein, & Pardo, 2012; see their Study 2 for the one exception). Baron et al. (2014) furthermore concluded that CRT items were more similar to math items without intuitive answers than they were to non-numeric verbal problems that had CRT-like intuitive answers.

Numeric ability itself has been associated with superior performance in a variety of judgment and decision tasks, making it plausible that numeracy may account for at least
part of the CRT’s association with better decision making. For example, Peters et al. (2006) found lower numeracy was related to more framing and format effects as well as denominator neglect. More numerate individuals, on the other hand, were less influenced by non-numerical information such as mood states and they demonstrated greater number-related affective reactions and sensitivity to different levels of numeric risk (Peters et al., 2009; see Peters, 2012; Reyna, Nelson, Han, & Dieckmann, 2009 for reviews). Numeracy effects are not limited to lab studies. McArdle, Smith, and Willis (2009) demonstrated that the more numerate accrue more wealth (even after accounting for demographic characteristics and other cognitive abilities, for example, working and long term memory), perhaps because the more numerate are less risk averse in their investments. We call the view that the CRT is primarily a measure of numeric ability and that numeric ability drives the CRT’s ability to predict better decisions, the Numeracy Hypothesis.

**Modeling Cognitive Reflection and Numeric Ability**

Researchers have begun to recognize that the processes underlying CRT performance may include both cognitive reflection and numeric ability (Böckenholt, 2012b; Campitelli & Gerrans, 2014; Del Missier, Mäntylä & Bruine de Bruin, 2012). Böckenholt (2012b) and Campitelli and Gerrans (2014), for example, assumed that solving a CRT problem required all participants initially to think of the incorrect intuitive response; then, their individual responses were determined in a two-step process of cognitive reflection and (if
cognitive reflection was high enough to detect the intuitive error) numeric ability. For example, the bat and ball problem brings to mind an intuitive response (10 cents). If cognitive reflection is high enough, a person checks the response and determines it is wrong ($1.10 + $0.10 \neq $1.10) and proceeds to the next step. To answer correctly (5 cents), the person may need the knowledge to set up the appropriate equation ($1.00 + x + x = $1.10); they may also need the capacity to solve the equation. If numeric ability is not high enough, an idiosyncratic non-intuitive error will emerge. In other CRT items, the person must be able to subtract, multiply and divide, and perhaps most important, know which operation is appropriate. Of course, some people may solve CRT problems through trial and error, simply trying various different options until they find one that satisfies the conditions of the problem.

This two-step process can be verified by recoding CRT responses into three categories (intuitive errors, non-intuitive errors, and non-intuitive correct responses) rather than the usual two categories of correct and incorrect. This additional information allows the separation of Cognitive Reflection (which categorizes intuitive responses from non-intuitive ones) from numeric ability (which separates non-intuitive correct responses from non-intuitive errors). Böckenholt (2012b) did this by treating cognitive reflection and numeric ability (labeled Inhibitory Control and Deliberate, respectively, in that paper) as separate latent variables in an item response theory model. This model fit better than a model with a single latent variable that was responsible for both checking the intuition and getting the correct answer (i.e., the simpler model effectively allowed only correct
and incorrect responses). He also showed an hypothesized diurnal effect on cognitive reflection vs. numeric ability. In particular, morning people showed greater cognitive reflection in the morning than the evening whereas evening people showed the opposite pattern (see also Bodenhausen, 1990). According to the author, no diurnal effect existed on the more trait-like (and presumably stable) numeric ability.

Campitelli and Gerrans (2014) produced a similar mathematical model and found that more cognitive reflection (labeled inhibition) was correlated with a greater likelihood to check intuitions in another cognitive bias: belief bias in syllogistic reasoning (Evans, Barston, & Pollard, 1983); their numeric ability construct (labeled mathematical computation) correlated with a three-item numeracy scale. However, they tested neither whether numeracy correlated with cognitive reflection nor whether belief bias correlated with numeric ability.

Although terminology and exact mathematical definitions of variables varied between the two studies, both studies conceptualized CRT responses as being comprised of cognitive reflection and numeric ability. In particular, Cognitive Reflection was the likelihood to give any non-intuitive answer and numeric ability was the conditional likelihood of giving the correct answer given that the answer was not intuitive.

Do Cognitive Reflection and Numeracy Both Predict Good Decision Making?

Although studies have demonstrated correlations of the CRT with decision biases, it is unclear whether the effects are due to cognitive reflection (as usually posited) or numeric
ability. Studies that separate cognitive reflection and numeric ability have not examined which is responsible for the CRT’s relations with decision-making biases and outcomes.

Two opposing hypotheses exist:

1. The Cognitive Reflection Hypothesis: Cognitive reflection will be responsible for the CRT’s correlations with decision-making abilities. Numeric ability will not account for this relation.

   However, cognitive reflection may only be predictive of decision making inasmuch as it correlates with numeric ability. In the present studies, we also examined performance of the Weller et al. (2013) numeracy scale. Because numeric ability may be a multi-faceted construct (Liberali et al., 2012; Weller et al., 2013) and the numeric skills required to solve CRT items are different from those tested on most numeracy scales, it is possible that the two numeric ability scales will account for different aspects of decision performance.

2. The Numeracy Hypothesis: Numeric ability measured on a numeracy scale and/or the CRT will account for the effects of cognitive reflection.

To test these hypotheses, we examined decision-making competence in two studies. To do so, we first used participants’ CRT responses to identify separate constructs of Cognitive Reflection and numeric ability through cognitive modeling and/or the proportions of responses falling into the three categories described above (intuitive errors, non-intuitive errors, and non-intuitive correct responses). We then examined the relations of these constructs of Cognitive Reflection and numeric ability with good decision
making. In Studies 1 and 2, we predicted consistency in risk perception from Bruine de Bruine, Parker, and Fischhoff’s (2007) Adult Decision Making Competence (ADMC) scale. In Study 2, we also examined relations with under/overconfidence (Bruine de Bruin et al., 2007), performance on incentivized risky gambles and intertemporal preferences (Frederick, 2005), and self-reported financial outcomes. In both studies, we considered whether a standard numeracy scale could account for any findings and used large, diverse samples. We focused on testing whether the Cognitive Reflection Hypothesis or the Numeracy Hypothesis provided the best explanation of the data.

Study 1

According to the Cognitive Reflection Hypothesis, greater cognitive reflection allows people to check faulty intuitions and, thus, reduce decision biases. Alternatively, the Numeracy Hypothesis posits that a lack of numeric ability produces these same biases. In the present study, we tested whether CRT performance was a significant predictor of decision biases due to Cognitive Reflection or numeric ability (called Calculation from here on when it is estimated from CRT responses). In Study 1, we attempted to find a bias that might be better predicted by Calculation rather than Cognitive Reflection. Consistent with the Numeracy Hypothesis, Del Missier et al. (2012) had found that consistency in risk perception (Bruine de Bruin et al., 2007) was predicted by numeracy, but not performance on inhibition tasks like the Stroop test. Although they did not test whether the CRT per se was predictive of consistency in risk perception, they did find
that numeracy and inhibition independently predicted scores on the CRT (Del Missier et al., 2012). Consistency in risk perception was therefore a good candidate task.

**Method**

**Participants and Procedure.**

As part of the Understanding America Study, data were collected over the internet from a diverse sample (N=1,413) from 5/31/14 to 10/22/14. Data collection was approved by the Institutional Review board of the University of Southern California. An address-based sampling method was used to recruit participants. Participants completed financial literacy questions, personality questions, the risk consistency subscale of the ADMC, and, finally, numeracy. Financial literacy and personality will not be discussed in the present paper. Participants were paid $10 to complete the survey which took, on average, about half an hour.

**Materials**

**Consistency in risk perception.** In the consistency in risk perception subscale of the ADMC, participants were asked to estimate the likelihood of a number of events (e.g., getting in a car accident) happening to them in the next year on a scale of 0% to 100%. The events are set up in such a way that participants can commit framing inconsistencies as well as conjunction inconsistencies for subset/superset relations and time (see below). Note that in the present study, we separated the three types of risk consistency scores
because they correlated only modestly and are predicted by different variables (especially the time conjunction score) as described below.

*Framing inconsistency.* Some of the events were complementary. The framing inconsistency score was the number of pairs of complementary items (out of four possible pairs) on which the sum of provided likelihoods was 10 or more points away from 100 (we introduced this threshold in order not to penalize participants who used more precise values; results were similar with other thresholds, including 5, 15 and 20; the 10 threshold worked best for the Cognitive Reflection Hypothesis and was retained). For example, if a participant indicated that his likelihood to drive accident free for the next 5 years was 80% and his likelihood to get into an accident in the next 5 years was 40%, then he would be scored as inconsistent for this pair of items.

*Subset/superset and time conjunction fallacies.* Some events were subsets of other events, for example, going to the dentist to fill a cavity was a subset of going to the dentist for any reason. The first conjunction fallacy (subset/superset) score was the number of times a subset event was judged as more likely than a superset event (out of four possible pairs). For example, if a participant indicated that her chance to go to a dentist in the next 5 years for any reason was 60%, and her chance to go to a dentist in the next 5 years to fill a cavity was 70%, then she would be scored as inconsistent for this pair of items.

The second conjunction fallacy (time) score was the number of times an event happening in the next year was judged as more likely than the same event happening in the next 5 years (out of 8 possible pairs). For example, if a participant indicated that his chance to...
go to the dentist in the next 5 years was 60% and his chance to go to the dentist in the next year was 70%, then he would be scored as inconsistent for this pair of items.

**Numeracy and CRT.** Participants completed the 8-item Rasch-based numeracy scale (Weller et al., 2013), which includes two CRT items. Participants also completed three additional CRT items (Toplak et al., 2014). Numeracy was scored as the proportion of non-CRT numeracy items answered correctly (out of a possible six items). Numeracy was mean-centered and standardized to match the scales of Cognitive Reflection and Calculation, which were estimated and scored as latent variables as described below, and as proportions. Cognitive Reflection was calculated as the proportion of CRT responses that were not the intuitive response (but they could be correct or incorrect; $\alpha=.48$). Calculation was computed as the proportion of non-intuitive CRT responses that were correct (i.e., it is the conditional probability of answering correctly *given that* the participant answered non-intuitively).

**Analyses**

We estimated a model identical to Böckenholt’s (2012b) Cognitive Miser model of the CRT. Their approach (unlike that of Campitelli & Gerrans, 2014) allows the estimation of individual differences and differences between items, accounts for measurement error, and allows the two abilities to be correlated. It is theoretically grounded in the Item Response Theory tradition. We used the nlme package Version 3.1 for linear and non-linear mixed-effects models (Pinheiro, Bates, DebRoy & Sarkar, 2014) to fit Böckenholt’s model because it handles missing observations and allows for dichotomous
response variables. De Boeck and Partchev (2012) described in detail how a package for
generalized linear mixed-effects models can be used to fit an IRTree model, of which the
Cognitive Miser model is one example (see also Böckenholt, 2012a). We describe this
method briefly below.

Responses to the five CRT items were treated as up to 10 repeated measures because it is
assumed that participants complete a two-step process when answering a CRT problem.
To respond correctly, they must successfully complete both steps. In Step A, they attempt
to avoid the intuitive response; if they fail to avoid it, processing is terminated and the
incorrect intuitive response is given. If they avoid the intuitive response, then participants
proceed to Step B and determine a non-intuitive response. If a participant reported the
incorrect intuitive response, the process was assumed to have terminated in Step A. Thus,
Step B was never performed, and the Step B response was treated as missing data. See
Table 3 for a depiction of how data were coded. We used model comparisons to test an
hypothesis concerning whether two separate abilities (vs. a single ability) were
responsible for completing steps A and B.

Table 3. Coding of Possible Responses

<table>
<thead>
<tr>
<th>Response</th>
<th>Step 1</th>
<th>Step 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intuitive Error</td>
<td>0</td>
<td>Missing</td>
</tr>
<tr>
<td>Non-intuitive Correct</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Non-Intuitive Error</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
**Model 1.** In the first model, we allowed only one factor to be responsible for individual differences in answering correctly on each step of each question, but estimated the population level difficulty for each step of each question. Based on this model’s constraint of having only one factor responsible for individual differences, if subject 1 is twice as likely to be correct on step A for a problem as subject 2, she must be twice as likely as subject 2 to be correct on that same problem’s step B, the next problem’s step A, etc. However, although the model constrains individual differences, it does allow step A in one problem to be more or less difficult than step B for the same problem which can be more or less difficult than step A for the next problem, etc. Hence, this model has 10 fixed effects: five coefficients for the difficulty of step A (one for each of five problems) and five coefficients for the difficulty of step B. In addition, it has one source of variation between people (random effect).

**Model 2.** As in Model 1, population-level differences in difficulty still exist between all the repeated measures. However, in Model 2, we allowed two abilities to explain sources of individual differences - one for step A (Cognitive Reflection) and the other for step B (Calculation). In this model, if subject 1 is twice as likely to be correct on step A of the first problem as subject 2, he is not necessarily twice as likely to be correct as subject 2 on step B of the same problem, but is still twice as likely to be correct as subject 2 on step A of the next problem. The correlation between Cognitive Reflection and Calculation was estimated; hence, Step A performance may or may not influence performance on Step B. This model has the same 10 fixed effects as Model 1, but it has two individual
difference parameters: (σ Cognitive Reflection, σ Calculation), and one parameter for the correlation between these abilities (γ). If this model fits better than the first model, we can conclude that two separate abilities influence CRT responses.

Results

**Identifying inattentive participants**

Inattentive participants would be counted as high on Cognitive Reflection because their nonsensical CRT responses would count as non-intuitive. We found and excluded four participants whose numeracy responses displayed a nonsensical pattern (e.g., entering 10 or 100 for most questions). Removing these participants did not substantially alter the results (results including these participants are available from the first author).

These exclusion criteria could be considered conservative, meaning that some inattentive participants may have given responses that did not exhibit a clear pattern. To allow for this possibility, we conducted robust regressions (available in Appendix A). These results mirror the results reported in the main text, but account for the possibility that a relatively small portion of the sample may score high on Cognitive Reflection and have large decision biases whereas the trend in the rest of the sample is the opposite. The similarity of these robust regressions to the results reported in the main text makes it unlikely that a relatively small group of inattentive participants influenced our results.
Descriptive statistics

The median participant earned between $50,000 and $60,000, was 49 years old, and had an associate degree; 52% of participants were female. See Table 4 for the proportion of participants giving each type of response on each item and Table 5 for means, standard deviations, correlations, and reliabilities of all scales.
Table 4. Intuitive and Correct Responses for CRT Items Used in Study 1

<table>
<thead>
<tr>
<th>Problem</th>
<th>Responses</th>
<th>Proportion of Responses That Are:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Int.</td>
<td>Correct</td>
</tr>
<tr>
<td>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? (in cents)</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>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?</td>
<td>24</td>
<td>47</td>
</tr>
<tr>
<td>Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class?</td>
<td>15, 30*</td>
<td>29</td>
</tr>
<tr>
<td>A man buys a pig for $60, sells it for $70, buys it back for $80, and sells it finally for $90. How much has he made?</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Simon decided to invest $8,000 in the stock market one day early in 2008. Six months after he invested, on July 17, the stocks he had purchased were down 50%. Fortunately for Simon, from July 17 to October 17, the stocks he had purchased went up 75%. At this point, Simon has: a. broken even in the stock market, b. is ahead of where he began, c. has lost moneys</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>57.0</td>
</tr>
</tbody>
</table>
Note. *The class grades question has two possible intuitive errors (15 and 30), both of which are quite common. Results are similar if one or both errors are counted as the intuitive error; both errors were counted as intuitive errors for purposes of the present paper.

Table 5. Correlations of the Measures in Study 1

<table>
<thead>
<tr>
<th></th>
<th>6-item Numeracy</th>
<th>Cognitive Reflection</th>
<th>Calculation</th>
<th>Frame Inconsistency</th>
<th>Conj. (Sets)</th>
<th>Conj. (Time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive Reflection</td>
<td>.46</td>
<td>.57</td>
<td>.67</td>
<td>-.24</td>
<td>-.23</td>
<td>-.23</td>
</tr>
<tr>
<td>Calculation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-.23</td>
<td>-.23</td>
</tr>
<tr>
<td>Frame Inconsistency</td>
<td>-.24</td>
<td>-.20</td>
<td>-.25</td>
<td></td>
<td>-.23</td>
<td>-.23</td>
</tr>
<tr>
<td>Conjunction (Sets)</td>
<td>-.23</td>
<td>-.09</td>
<td>-.15</td>
<td></td>
<td>-.08</td>
<td>-.08</td>
</tr>
<tr>
<td>Conjunction (Time)</td>
<td>-.05</td>
<td>-.03</td>
<td>-.08</td>
<td></td>
<td>.20</td>
<td>.17</td>
</tr>
<tr>
<td>Mean</td>
<td>0</td>
<td>-0.01</td>
<td>0.04</td>
<td>1.27</td>
<td>0.40</td>
<td>1.83</td>
</tr>
<tr>
<td>SD</td>
<td>1</td>
<td>0.86</td>
<td>1.27</td>
<td>1.10</td>
<td>0.66</td>
<td>0.86</td>
</tr>
<tr>
<td>Reliability (alpha)</td>
<td>.67</td>
<td>.54</td>
<td>--</td>
<td>.43</td>
<td>.04</td>
<td>.37</td>
</tr>
</tbody>
</table>

Note. All correlations were significant at the .05 level except Conjunction (Time) with numeracy (p=.07) and Cognitive Reflection (p=.37). The alpha for Cognitive Reflection represents the unstandardized Cronbach’s alpha for the number of items that were answered with any non-intuitive response. Alpha for Calculation cannot be calculated because this variable is either a latent variable (Study 1) or a proportion with a variable denominator (Studies 1 and 2).
Modeling

Model 2 fit the data substantially better than Model 1 (change in BIC = 198; χ² (2) = 217, p<.001), replicating previous results (Böckenholt, 2012b; Campitelli & Gerrans, 2014). Model 1 results consisted of the same fixed effects as in Model 2, but estimated less accurately. Therefore, we report only the results of Model 2.

Calculation varied more in the sample (σ Calculation = 2.1) than did Cognitive Reflection (σ Cognitive Reflection = 1.2), suggestive of the CRT measuring individual differences in Calculation to a larger degree than Cognitive Reflection. This is similar to Calculation having a higher reliability than Cognitive Reflection in a traditional analysis (Cronbach’s alpha for Calculation cannot be calculated because this variable is either a latent variable or a proportion with a variable denominator). The two abilities correlated substantially (γ=.40) similar to γ=.31, calculated from the variances and covariances provided by Böckenholt (2012b); a correlation could not be computed for Campitelli and Gerrans’ model as it does not estimate variances or covariances. We also found substantial differences in difficulty in both Cognitive Reflection and Calculation among the items (see Table 6). Coefficients in the table indicate log odds, so a coefficient of 0 indicates that participants were, on average, as likely to do the task correctly as they were to fail; higher coefficients indicate greater chances of doing the task correctly (e.g., 0.3 indicates the odds of answering correctly vs. incorrectly are e⁰.³ = 1.35, and the probability of answering correctly is .57). Consistent with Frederick (2005), Calculation was easier than Cognitive Reflection; however, Calculation was far from trivial. For example, in the Bat
and Ball problem, Calculation ($\beta=0.32$) was substantially easier than Cognitive Reflection ($\beta=-1.64$); however, people still failed to calculate correctly almost half the time. Calculation in the investment problem, on the other hand, was quite easy ($\beta=2.10$). This is sensible because the investment problem is multiple choice so that simply eliminating the intuitive option narrows the set of choices to only two possibilities.

Table 6. Model difficulty parameters (standard errors) for each CRT item.

<table>
<thead>
<tr>
<th>Item</th>
<th>Cognitive Reflection</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bat and ball</td>
<td>-1.65 (0.08)</td>
<td>0.32 (0.18)</td>
</tr>
<tr>
<td>Lily pad</td>
<td>-0.23 (0.07)</td>
<td>0.23 (0.12)</td>
</tr>
<tr>
<td>Class size</td>
<td>-0.89 (0.07)</td>
<td>-0.28 (0.14)</td>
</tr>
<tr>
<td>Pig sale</td>
<td>0.38 (0.07)</td>
<td>0.03 (0.11)</td>
</tr>
<tr>
<td>Investment</td>
<td>0.37 (0.07)</td>
<td>2.10 (0.14)</td>
</tr>
</tbody>
</table>

To determine how Cognitive Reflection and Calculation related to numeracy and decision performance, we estimated the random effects of these variables by participant. These random effects are the modes of the distributions of Cognitive Reflection and Calculation conditional on the model for each participant (i.e., the most likely Cognitive Reflection and Calculation scores given that Model 2 is correct). In other words, Cognitive Reflection and Calculation, as discussed below, are scores for these constructs for each participant derived from the model. As expected, we found that greater numeracy was correlated with greater Calculation ($r = .57, p < .001$), replicating Campitelli and Gerrans’ (2014) finding. However, we also found that Cognitive Reflection had roughly the same
correlation with numeracy ($r = .46$, $p < .001$). Greater Calculation was also correlated with greater Cognitive Reflection ($r = .67$, $p < .001$); the correlation explicitly estimated in the model, $\gamma = .40$, is likely more reliable. This correlation may reflect a general ability like intelligence, or something more specific to performance on CRT problems.

Since Cognitive Reflection and Calculation were correlated, and each was substantially correlated with numeracy, we conducted multiple regressions for each decision bias to partial out shared variance and, hence, to test which part of the CRT independently predicted biases (see Table 7). We examined the participants who completed all the tasks: consistency in risk perception subscale, numeracy, and CRT (final $N = 1,225$). Analyses conducted on all participants who completed each subscale produced similar results as did robust regression analyses (see Appendix A) and multiple regressions using Cognitive Reflection and Calculation scores computed as proportions.²

² These estimates are subject to measurement error and, in the case of Calculation, are heteroskedastic. Nonetheless, estimates were highly correlated to the latent scores (.99 for Cognitive Reflection, .88 for Calculation) and produced similar regression results.
Table 7. Regression Analyses - Consistency in Risk Perception and CRT

<table>
<thead>
<tr>
<th></th>
<th>Frame Inconsistency</th>
<th>Conjunction (subset vs. superset)</th>
<th>Conjunction (time)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without Numeracy</td>
<td>With Numeracy</td>
<td>Without Numeracy</td>
</tr>
<tr>
<td>Intercept</td>
<td>2.14 (0.20)</td>
<td>2.04 (0.20)</td>
<td>0.98 (0.12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.11 (0.23)</td>
</tr>
<tr>
<td>Cognitive Reflection</td>
<td>-0.07 (0.05)</td>
<td>-0.06 (0.05)</td>
<td>0.04 (0.03)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.08 (0.06)</td>
</tr>
<tr>
<td>Calculation</td>
<td>-0.12 (0.03)</td>
<td>-0.09 (0.03)</td>
<td>-0.05 (0.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-0.10 (0.04)</td>
</tr>
<tr>
<td>Numeracy</td>
<td>-- (0.04)</td>
<td>-0.13 (0.04)</td>
<td>-- (0.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.00 (0.05)</td>
</tr>
<tr>
<td>F</td>
<td>20.4</td>
<td>19.1</td>
<td>12.7</td>
</tr>
<tr>
<td>df</td>
<td>6, 1218</td>
<td>7, 1217</td>
<td>6, 1218</td>
</tr>
<tr>
<td>R²</td>
<td>.09</td>
<td>.09</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.01</td>
</tr>
</tbody>
</table>

Note. Each dependent variable was regressed onto Cognitive Reflection and Calculation and all demographic variables (age, income, education and gender), though their coefficients were not reported for simplicity. The results were reported in the columns titled “Without Numeracy.” Numeracy was then added and the results were reported in the columns titled “With Numeracy.” Values are unstandardized beta coefficients with standard errors in parentheses. Bold font denotes statistical significance at p<.05.

**Frame Inconsistency**

In multiple regression with frame inconsistency as the dependent variable, we found that higher Calculation (p<.001) but not Cognitive Reflection (p=.15) independently predicted less frame inconsistency (e.g., more consistency between estimated likelihoods to drive accident free versus get into an accident) after accounting for demographic variables. When numeracy was added to the model (p=.001), it added significant independent
predictive power and did not completely account for the variance explained by Calculation (p=.01).

**Conjunction Fallacies**

Greater Calculation (p=.01), but not Cognitive Reflection (p=.13), predicted fewer conjunction fallacies between subset and superset events (e.g., participants estimated more consistent likelihoods between going to the dentist for any reason and going to the dentist to fill a cavity). When numeracy was added to the model (p<.001), it accounted for the variance explained by Calculation (p=.33). Calculation (p=.01) also predicted conjunction fallacies between points in time; Cognitive Reflection did not (p=.16). When numeracy was added to the model, it did not explain additional significant variance (p=.93), and it did not account for the effects of Calculation (p=.01). We examine time conjunction fallacies again in Study 2.

**Discussion**

The present study replicated and extended earlier results from Campitelli and Gerrans (2014). In particular, Cognitive Reflection and Calculation behaved like distinct abilities, and Calculation was positively correlated with numeracy. However, Cognitive Reflection was positively correlated with numeracy as well; this correlation had not been tested in earlier studies. This finding, however, may not be surprising given that a numeric formula is needed to check the intuition in CRT problems (e.g., in the bat and ball problem, $1.00 + $0.10 + $0.10 ≠ $1.10). Thus, numeracy may be important to both steps in
solving CRT problems; setting up a numeric formula is necessary to check intuitions and adequate numeric ability is necessary to solve the formula.

Contrary to the Cognitive Reflection Hypothesis, Cognitive Reflection did not provide any unique explanatory power in Study 1’s decision tasks whereas Calculation and numeracy did in both tasks. It was not entirely clear whether the non-significance of Cognitive Reflection in predictions of conjunction fallacies may have been due to numeric ability accounting for its effects or because it was not a potent predictor in the first place. Our model also showed that the CRT measures Calculation to a greater degree than it measures Cognitive Reflection. Therefore, our results could be explained in part by Calculation’s relatively low reliability (though its reliability was not much lower than that of numeracy). This is important because it suggests that previous results that attribute the predictiveness of the CRT to a cognitive reflection construct may be in error, given that the scale measures Calculation to a greater degree. Our results were most consistent with the Numeracy Hypothesis although we had not expected Calculation to be predictive beyond numeracy. We offer a possible explanation in the general discussion. The dependent measures in this study were derived from a subscale of decision making competence, consistency in risk perception, which we expected to correlate with Cognitive Reflection but be explained by numeric ability (Bruine de Bruin et al., 2007; Del Missier et al., 2012).

Note that conjunction fallacies regarding subsets and supersets had unacceptable reliability, as measured by Cronbach’s alpha, even compared to the relatively low
reliability of the other decision biases. However, these conjunction errors correlated substantially with our predictors and the framing bias. The fact that the reliability was lower than the variance explained in our models (this was still true if adjusted R² was used instead) suggests that either our results were due to chance or that Cronbach’s alpha measure of reliability may not be appropriate, perhaps because it is an estimate of the lower bound of reliability (Cronbach & Shavelson, 2004). Indeed, other measures of reliability (Sijtsma, 2009) gave considerably higher estimates, but we used alpha because it is the most common estimate and is often used for binary data like ours. These low reliabilities point to the need to replicate the present results. In Study 2, we attempted to replicate our results, but also turned to tasks that have been related to CRT performance more traditionally in past research. We also examined directly incentivized tasks and more real-world decision outcomes.

**Study 2**

In Study 2, we again examined the ADMC’s consistency in risk perception, but we also focused on decision tasks more traditionally associated with cognitive reflection. In particular, we examined under/overconfidence (another subscale of the ADMC). Hoppe and Kusterer (2011) demonstrated that correct levels of confidence were correlated with higher CRT scores (see Del Missier et al., 2012 for similar results with a presumably related inhibition measure). However, other research suggests that numeracy may be independently predictive as well (Winman, Juslin, Lindskog, Nilsson, & Kerimi, 2014). We also examined intertemporal and risky choices similar to those originally studied by
Frederick (2005). The Cognitive Reflection Hypothesis suggests that the CRT’s predictive ability in these tasks is due to cognitive reflection, not numeric ability. Consistent with the Numeracy Hypothesis, however, research has demonstrated that greater numeracy is related both to morepatience in intertemporal choice and more expected-value-consistent risky choices (Benjamin, Brown, & Shapiro, 2013). In addition, we examined whether CRT and/or numeracy would be associated with inconsistent responses in risky choices. In particular, we expected that lower numeracy or worse CRT performance would be associated with risky choices that were logically inconsistent with previously expressed preferences. No previous studies have considered CRT or numeracy relations with this inconsistency.

Finally, we examined self-reported financial outcomes and predicted that both cognitive reflection and numeric ability would independently predict having retirement savings, paying bills on time, and not taking predatory loans. Avoiding undesirable financial outcomes likely requires understanding how costly bad financial moves can be; less numerate individuals do not fare well in this regard (Soll, Keeney, & Larrick, 2013). It also may require self-regulation (related to cognitive reflection by Böckenholt, 2012) to control impulsive spending (Vohs & Faber, 2003). Thus, we expected that both numeric ability and Cognitive Reflection would independently predict positive financial outcomes.
Method

Procedure

Participants in RAND’s American Life Panel (ALP: www.rand.org/labor/roybalfd/american_life.html) were paid $20 to complete each half hour Internet survey. Data collection was conducted and approved by RAND Corporation. The various questionnaires described below were administered at different points in time. A total of 1,478 participants provided demographic information and responses to CRT and other numeracy items. Stepwise regressions to predict the decision bias composite were conducted on the 939 participants who completed those items and at least one each of the intelligence measures and decision-bias tasks. Stepwise regressions to predict the financial outcome composite were conducted on the 1,131 participants who completed demographics, numeracy, CRT, and at least one each of the intelligence and financial-outcome tasks.

Measures

We examined the same ADMC subscales as in Study 1 and several additional decision making tasks. Participants also completed the Weller et al. (2013) numeracy scale and an additional CRT item.
Consistency in risk perception

A complete version of the consistency in risk perception scale was administered and scored in the same way as in Study 1. The scale included four pairs of framing inconsistency pairs, six subset/superset conjunction pairs, and 10 time conjunction pairs.

Under/over/accurate confidence. Participants were asked if they thought fourteen general knowledge statements (e.g., “Amman is the capital of Jordan”) were true or false, and they indicated their confidence that they answered each item correctly from 50% (just guessing) to 100% (absolutely sure). We used the absolute difference between the percentage of items answered correctly and the average confidence across items to assess confidence accuracy.

Incentivized intertemporal choice. Participants were asked if they wanted their payment for the survey to be mailed immediately, or 110% of those payments to be mailed 10 days later. Participants were shown the amounts they would be mailed in each case and were rewarded according to the plans they chose. This variable was coded 0 (indicating a preference for more money later) or 1 (indicating a preference for less money now).

Incentivized risky choice. The Holt-Laury Procedure was employed (Holt & Laury, 2002). Specifically, participants were asked their preferences between ten pairs of gambles, all in the domain of gains. Each pair of gambles included one safe gamble, in which the participant could win either $2.00 with some probability, otherwise get $1.60, and one risky gamble, in which the participant could win $3.85 with the same probability, otherwise get $0.10. The response consistent with an expected-value calculation was the
safe gamble when the probability was low (e.g., the gamble “10% chance to win $2.00, otherwise $1.60” has a higher expected value than the gamble “10% chance to win $3.85, otherwise $0.10”) and the risky gamble when the probability was high (e.g., “90% chance to win $2.00, otherwise $1.60” has a lower expected value than “90% chance to win $3.85, otherwise $0.10”). In the case of 100%, the “risky gamble” amounted to receiving $3.85 and dominated the “safe gamble,” which amounted to receiving $2.00. Each participant was given a risk aversion score, which was the number of times the safe gamble was chosen. One of the choices was played for real, and any payoff was added to the participant’s survey payment.

Participants chose between the two gambles at 10 probabilities in the same fixed order (10%, 20%, 30%,…100%). Due to the fixed order, once the risky gamble was chosen, it should be preferred in all subsequent choices regardless of risk preferences because increasing the probability simply makes it better compared to the safe gamble. Hence, each participant was also given a consistency score, which was the minimum number of choices that had to be changed so that the participant would have consistent preferences.

**Numeracy and CRT**

Three items are too few for the latent variable modeling approach of Study 1, and the model Campitelli and Gerrans (2014) used was inappropriate because it is unable to produce individual scores for participants. Instead, Cognitive Reflection and Calculation were estimated using proportions (see Study 1). We chose this approach because it is conceptually similar to and more transparent than earlier modeling approaches.
(Böckenholt, 2012b; Campitelli & Gerrans, 2014). About 24% of participants answered all CRT problems intuitively, making their Calculation scores nonsensical\(^3\). We gave these participants Calculation scores of 0 although results were essentially identical if scores were instead imputed using linear fits from variables correlated with Calculation including numeracy, education, income and gender (Enders, 2010).\(^4\) The same 8-item numeracy scale as in Study 1 was administered, and 6 of its non-CRT items were used as a measure of numeracy (\(\alpha = .58\)).

**Financial outcomes**

Participants reported five financial decision-making outcomes (see Table 8). Each was coded 0 if the outcome was unfavorable and 1 if favorable.

\(^3\) The conditional probability of answering correctly given that the participant answered non-intuitively cannot be calculated if the participant never answered non-intuitively.

\(^4\) In Study 1, the correlations of Calculation with the latent variable and proportion score were identical if Calculation was scored 0, or if participants who answered all problems intuitively were omitted. This result indicates that this imputation technique is consistent with how the latent variable model treated these participants.
Table 8. Financial Outcomes

<table>
<thead>
<tr>
<th>Name</th>
<th>Question</th>
<th>Answers counted as “success”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avoided predatory loans</td>
<td>Within the last year, have you obtained credit from a rent-to-own store, pawn shop, payday lender, cash advance lender, auto title lender, or tax return preparer?</td>
<td>No</td>
</tr>
<tr>
<td>Avoided being denied credit</td>
<td>Have you been denied credit for any type of loan within the last year?</td>
<td>No</td>
</tr>
<tr>
<td>Saved money for retirement</td>
<td>What is the total amount of wealth you have accumulated so far for the purpose of retirement preparation, including both accounts like 401k or IRA and also any other types of accounts or forms of retirement saving?</td>
<td>Not 0</td>
</tr>
<tr>
<td>Loans on time</td>
<td>Have you made a late payment on any loan in the last year?</td>
<td>No</td>
</tr>
<tr>
<td>Paid credit cards in full*</td>
<td>Over the past 12 months, I always paid my credit cards in full.</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note. *Only participants who said they had a credit card in the past 12 months (N=1,207) were asked about whether they paid it in full.

**Intelligence measures**

Participants completed four non-numeric intelligence measures (Raven’s Matrices, antonyms, a vocabulary measure that required identification of words from pictures, and
verbal analogies). Scores indicated the number of questions answered correctly.\(^5\) Scores on each test were standardized (i.e., divided by its standard deviation and mean centered) and averaged to derive a composite intelligence measure (standardized \(\alpha=.48\)). If a score on a particular test was missing for a participant, only scores on the remaining tests were used to calculate that participant’s composite score\(^6\).

**Composites**

To avoid testing our seven decision biases and five financial outcomes one at a time, we created two composites\(^7\) (see Toplak et al., 2011 for the use of composites in a similar context). A decision-bias composite was computed as the average of the standardized decision-bias variables (Framing Inconsistency, Conjunction (subset vs. superset), and Conjunction (time), Under/Overconfidence, Impatient Intertemporal Choice, Risk Aversion and Risk Inconsistency). Standardized alpha was low (.37), but comparable to

\[\]

\(^5\) Participants also completed sequences of numbers, but these scores were omitted due to shared variance and construct overlap with numeracy.

\(^6\) Results were nearly identical if Raven’s Matrices (as a proxy of fluid intelligence) and vocabulary (as a proxy of crystallized intelligence) were both included as predictors, rather than including the composite, or if all four measures were included separately.

\(^7\) Testing each decision outcome and each financial outcome one at a time while controlling for intelligence was also difficult because excluding participants who did not complete the intelligence measure substantially decreased the sample size for a number of the outcomes. In addition, conducting twelve stepwise regressions would result in excessive Type 1 error rates due to the multiple tests (stepwise regression may result in overfitting, even when conducted just once; Babyak, 2004). Finally, by averaging decision biases, we reduced error. Nonetheless, regressions were also conducted separately for each individual decision bias with similar results that are available from the first author.
previous research (e.g., Toplak et al., 2011). Scores were standardized and averaged in a manner similar to the intelligence measures above. Thus, if a participant only completed the under/overconfidence measure, and scored 1 standard deviation (SD) higher than the mean, 1 would serve as his/her decision-bias score. But if that participant also completed the risk-preference choices and scored 1 standard deviation lower than the mean on both risk aversion and choice inconsistencies, he would receive a score of $(1-1-1)/3 = -1/3$. A financial-outcome composite was computed as the number of positive financial outcomes divided by the number of financial-outcome questions answered.

**Results**

*Identifying inattentive participants.*

Eighteen (out of 1,478) participants were deleted due to numeracy responses that followed a pattern indicating inattention (e.g., entering 10 or 100 for most questions). Their deletion did not significantly alter results.

*Replications.*

As in Study 1, greater Cognitive Reflection was correlated with greater Calculation and both were correlated with higher numeracy (Table 9). We replicated Study 1’s framing inconsistency results: Greater Cognitive Reflection was correlated with less bias ($r= - .12$, $p < .001$). In multiple regression and after controlling for demographic variables, however, Calculation and numeracy accounted for the effects of Cognitive Reflection (regression coefficients are Calculation: $b= - 0.05$, $p= .07$, numeracy: $b= - 0.16$, $p= .002$, and Cognitive
Reflection: $b=-0.01; p=.81$, final model $F_{7,918}=6.07; p<.001; R^2=.04)$. We also found that greater Cognitive Reflection was correlated with showing fewer conjunction fallacies between subset and superset events ($r=-.10$, $p=.002$). Again, when Calculation and numeracy were added in multiple regression, they accounted for the effects of Cognitive reflection (regression coefficients are Calculation: $b=-0.06$, $p=.28$, numeracy: $b=-0.35$, $p=.004$, and Cognitive Reflection: $b=-0.004; p=.96$, final model $F_{7,880}=4.3; p<.001; R^2=.03)$. Only age and gender (and none of our cognitive predictors) were related to time conjunction fallacies.
Table 9. Correlations of independent measures among themselves and with DVs

| Independent Vars.          | N   | M   | Cognitive Reflection | Calculation | Numeracy | Intelligence | Education | Income | Age | Gender |                      |                      |                     |                      |                      |                      |                      |                      |
|---------------------------|-----|-----|----------------------|-------------|----------|--------------|-----------|--------|-----|--------|-----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
|                          |     |     |                      |             |          |              |           |        |     |        |                       |                       |                      |                       |                      |                      |                      |
| Cognitive Reflection      | 1459| 0.44|                      |             |          |              |           |        |     |        |                       |                       |                      |                       |                      |                      |                      |
| Calculation               | 1459| 0.47|                      |             |          |              |           |        |     |        |                       |                       |                      |                       |                      |                      |                      |
| Numeracy                  | 1459| 0.62|                      |             |          |              |           |        |     |        |                       |                       |                      |                       |                      |                      |                      |
| Intelligence              | 1135| 0.20|                      |             |          |              |           |        |     |        |                       |                       |                      |                       |                      |                      |                      |
|                            |     | 0.61|                      | 0.44        | 0.51     |              | 0.24      | 0.26    | 0.37 | 0.23  |                       |                       |                      |                       |                      |                      |
| Education                 | 1460| 11.7|                      |             |          |              |           |        |     |        |                       |                       |                      |                       |                      |                      |                      |
|                            |     |     |                       | 0.19        | 0.25     | 0.30         | 0.24      | 0.23   | 0.22 | 0.41  |                       |                       |                      |                       |                      |                      |
| Income                    | 1460| 57  |                      |             |          |              |           |        |     |        |                       |                       |                      |                       |                      |                      |                      |
|                            |     | 0.07|                      | -0.06       | -0.04    | -0.03        | 0.01      | 0.01   | 0.04 | 0.04  |                       |                       |                      |                       |                      |                      |
| Age                       | 1460| 0.56|                      | -0.21       | -0.23    | -0.24        | 0.01      | 0.01   | 0.04 | 0.04  |                       |                       |                      |                       |                      |                      |
| Gender                    | 1460|     |                       |             |          |              |           |        |     |        |                       |                       |                      |                       |                      |                      |                      |
|                            |     |     |                       |             |          |              |           |        |     |        |                       |                       |                      |                       |                      |                      |                      |
| Framing                   | 926 | 0.64|                      | -0.12       | -0.15    | -0.17        | -0.16     | -0.11  | -0.08 | -0.08 |                       |                       |                      |                       |                      |                      |
| Inconsistency             | 888 | 2.4 |                      | -0.10       | -0.04    | -0.08        | -0.05     | -0.03  | -0.00 | -0.00 |                       |                       |                      |                       |                      |                      |
| Conj. (set)               | 871 |     |                       |             | 0.02     | 0.01         | 0.01      | 0.01   | 0.01 | 0.01  |                       |                       |                      |                       |                      |                      |
| Conj. (time)              |     |     |                       |             |          |              |           |        |     |        |                       |                       |                      |                       |                      |                      |                      |
| Under/Overconfidence      | 913 | 7.8 |                      | -0.04       | -0.08    | -0.08        | -0.10     | -0.02  | -0.07 | -0.07 |                       |                       |                      |                       |                      |                      |
| Impatient Inter-          | 140 | 0.10|                      | -0.10       | -0.01    | -0.25        | -0.26     | -0.24  | -0.27 | -0.13 |                       |                       |                      |                       |                      |                      |
| temporal Choice           |     |     |                       |             |          |              |           |        |     |        |                       |                       |                      |                       |                      |                      |                      |
| Risk Averse Choices       | 1043| 5.1 |                      | -0.02       | -0.05    | -0.08        | -0.07     | -0.05  | -0.07 | -0.07 |                       |                       |                      |                       |                      |                      |
| Risky Choice Inconsistencies | 1043| 0.59|                      | -0.15       | -0.17    | -0.25        | -0.21     | -0.10  | -0.12 | -0.12 |                       |                       |                      |                       |                      |                      |
| Decision Bias Composite   | 1178|     |                       | -0.13       | -0.18    | -0.24        | -0.21     | -0.13  | -0.15 | -0.15 |                       |                       |                      |                       |                      |                      |
| Avoided Predatory Loans   | 1385| 0.97|                      | 0.07        | 0.09     | 0.09         | 0.07      | 0.12   | 0.14 | 0.14  |                       |                       |                      |                       |                      |                      |
| Avoided Being Denied Credit | 1385| 0.91|                      | 0.06        | 0.07     | 0.12         | 0.11      | 0.09   | 0.13 | 0.13  |                       |                       |                      |                       |                      |                      |
| Saved Money for Retirement| 465 | 0.87|                      | 0.16        | 0.22     | 0.18         | 0.17      | 0.27   | 0.36 | 0.36  |                       |                       |                      |                       |                      |                      |
| Loans on Time             | 1385| 0.86|                      | 0.03        | 0.02     | 0.02         | -0.01     | 0.03   | 0.03 | 0.03  |                       |                       |                      |                       |                      |                      |
| Paid Credit               | 1200| 0.47|                      | 0.14        | 0.10     | 0.16         | 0.13      | 0.15   | 0.10 | 0.10  |                       |                       |                      |                       |                      |                      |
| Financial Composite       | 1396| 0.82|                      | 0.11        | 0.11     | 0.14         | 0.12      | 0.14   | 0.13 | 0.13  |                       |                       |                      |                       |                      |                      |

78
New decision biases and financial outcomes.

Greater numeracy was related to showing less of each of the other decision biases whereas greater Cognitive Reflection was only significantly related with fewer risky-choice inconsistencies (see Table 9). In multiple regression of risky choice inconsistencies, however, numeracy accounted for the effects of Cognitive Reflection (coefficients for numeracy, Calculation, and Cognitive Reflection were b=-0.77, p<.001, b=-0.09, p=.20 and b=-0.05, p=.57 respectively; final model F3,1039=23.6; p<.001; R²=.06). Cognitive Reflection did not correlate with less risk aversion or more patient intertemporal choice (see Campitelli & Labollita, 2010 and Oechssler, Roider & Schmitz, 2009 for similar results). Campitelli and Labollita (2010) also found that Cognitive Reflection was related to more choices consistent with expected value. Similarly, we found that greater Cognitive Reflection correlated with more expected-value-consistent choices (r=.15, p<.001). However, when numeracy and Calculation were added as predictors, Cognitive Reflection (b=-0.02, p=.91) and Calculation (b=0.17, p=.27) became non-significant, whereas numeracy remained significant (b=2.36, p<.001, F3,1039=34.4, p<.001, R²=.09). Both Cognitive Reflection and numeracy correlated with each of the financial outcomes except making late loan payments.

As expected, we also found that greater intelligence, more education, greater income, younger age, and being male were correlated with greater Cognitive Reflection,
Calculation, and numeracy. These potentially confounding variables were also correlated with decision biases and financial outcomes, possibly explaining the effects reported above. Thus, we conducted stepwise regressions to determine whether Cognitive Reflection and/or numeric ability retained independent predictive power above and beyond these variables.

**Stepwise regressions.**

For both composites, we conducted a stepwise regression, adding variables in the following order: (1) gender, age, education, income, (2) intelligence composite, (3) Cognitive Reflection, (4) Calculation, and (5) numeracy. Full regression results are available in Table 10. In predictions of the decision-bias composite, demographic variables made little difference with the exception of greater income predicting fewer decision biases (model F_{4,929} = 4.9, p<.001, R^2=.02). Greater intelligence was related to less bias as expected (Stanovich & West, 1998; b= -0.21, p<.001; change in R^2=.02) and accounted for the effects of the demographic variables. Cognitive Reflection was a borderline significant predictor of decision biases beyond intelligence^8^ (b= -0.11, p=.052, change in R^2=.004). Greater Calculation was associated with fewer biases (b= -0.17, p<.001, change in R^2=.01) above and beyond IQ and Cognitive Reflection despite its

---

^8^ Cognitive Reflection was predictive if it was entered without intelligence in the model (b= -0.14, p=.008); however, its effects could still be completely accounted for by numeracy (b= -0.03, p=.60).
high correlation with the latter; Calculation completely accounted for the effects of Cognitive Reflection. Numeracy was also a significant predictor of fewer biases (b=-0.42, p<.001, change in $R^2=.02$); it did not fully account for the effects of Calculation, which remained significant after numeracy was added (b=-0.12, p=.02; full model $F(8,925)=10.5$, p<.001, $R^2=.07$).
Table 10. Stepwise regression results predicting decision-bias and financial-outcome composites

<table>
<thead>
<tr>
<th>Dep. Variable</th>
<th>Intercept</th>
<th>Education</th>
<th>Income</th>
<th>Age*</th>
<th>Gender</th>
<th>Intel</th>
<th>Cog Refl</th>
<th>Calculation</th>
<th>Numeracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Decision Biases</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td>0.17</td>
<td>-0.01</td>
<td><strong>-0.01</strong></td>
<td>0.01</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.01)</td>
<td>(0.005)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 2</td>
<td>0.06</td>
<td>-0.00</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.06</td>
<td><strong>-0.21</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 3</td>
<td>0.08</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.04</td>
<td><strong>-0.20</strong></td>
<td>-0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.06)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 4</td>
<td>0.09</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.03</td>
<td><strong>-0.19</strong></td>
<td>-0.02</td>
<td><strong>-0.17</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.05)</td>
<td></td>
</tr>
<tr>
<td>Step 5</td>
<td>0.24</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.01</td>
<td><strong>-0.15</strong></td>
<td>0.06</td>
<td><strong>-0.12</strong></td>
<td><strong>-0.42</strong></td>
</tr>
<tr>
<td></td>
<td>(0.15)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.04)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.05)</td>
<td>(0.11)</td>
</tr>
<tr>
<td><strong>Financial Outcomes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td><strong>0.57</strong></td>
<td><strong>0.007</strong></td>
<td><strong>0.005</strong></td>
<td><strong>0.027</strong></td>
<td><strong>-0.03</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.003)</td>
<td>(0.002)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 2</td>
<td>0.59</td>
<td>0.005</td>
<td>0.004</td>
<td>0.028</td>
<td><strong>-0.03</strong></td>
<td><strong>0.04</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.003)</td>
<td>(0.002)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 3</td>
<td>0.59</td>
<td>0.005</td>
<td><strong>0.004</strong></td>
<td><strong>0.028</strong></td>
<td><strong>-0.02</strong></td>
<td>0.04</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.003)</td>
<td>(0.002)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 4</td>
<td><strong>0.59</strong></td>
<td>0.004</td>
<td><strong>0.004</strong></td>
<td><strong>0.028</strong></td>
<td><strong>-0.02</strong></td>
<td>0.03</td>
<td>0.03</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.003)</td>
<td>(0.002)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td></td>
</tr>
<tr>
<td>Step 5</td>
<td><strong>0.55</strong></td>
<td>0.003</td>
<td>0.003</td>
<td><strong>0.029</strong></td>
<td><strong>-0.02</strong></td>
<td>0.03</td>
<td>0.01</td>
<td><strong>-0.01</strong></td>
<td><strong>0.10</strong></td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.03)</td>
</tr>
</tbody>
</table>

Note. Unstandardized beta coefficients with standard errors in parentheses. Cog Refl = Cognitive Reflection. Intel=intelligence. Numeracy is the six-item numeracy scale used in both Studies 1 and 2. Bold font denotes statistical significance. *Age was divided by 10 in this regression to make the coefficients more interpretable (effects of age are effects of being a decade older). Gender was coded 0 when male and 1 when female. We predicted the financial-outcomes composite using a similar approach. In Step 1, demographic variables were predictive (model F_{4,1118} = 18.6, R^2=.06), with more positive
financial outcomes among those with greater education, income, and age. Higher intelligence was also predictive of better outcomes (b=0.04, p=.002, change in $R^2=.01$). Cognitive Reflection was not a significant predictor of better financial outcomes (b=0.03, p=.10, change in $R^2=.002$) above and beyond demographics and intelligence, even before accounting for Calculation and numeracy. Calculation was also not significant (b=0.01, p=.74, change in $R^2<.001$). In the final model, greater numeracy did predict better financial outcomes (b=0.10, p=.004, change in $R^2=.01$) as did higher income and intelligence (full model F(8,1114)=12.1, p<.001, $R^2=.08$).

Discussion

As in Study 1, we found that numeric ability, not Cognitive Reflection, predicted framing inconsistencies and conjunction fallacies between subsets and supersets. We did not find any decision biases that Cognitive Reflection predicted independently. We also found that Calculation and numeracy, but not Cognitive Reflection, predicted a decision-bias composite that included subscales of the ADMC and the original two biases tested by Frederick (2005). In addition, only numeracy predicted financial outcomes independently. These results were inconsistent with the Cognitive Reflection Hypothesis.

9 Cognitive Reflection was predictive if it was entered without intelligence in the model (b=-0.04, p=.046); however, its effects could still be completely accounted for by numeracy (b=-0.01, p=.54).
and supported the Numeracy Hypothesis. Our findings cannot be explained by the high correlation between Cognitive Reflection and Calculation, since Cognitive Reflection was not predictive of either composite before Calculation was included in the model. Numeracy was also related to less under/overconfidence (see also Winman, Juslin, Lindskog, Nilsson, & Kerimi, 2014). This finding is reasonable given that under/overconfidence is a task in which participants are asked to produce probabilities. However, contrary to expectations, Cognitive Reflection was not associated with more accuracy in this task. It may be that Hoppe and Kusterer’s (2011) finding that greater CRT scores were related to more confidence accuracy was due to them not separating Cognitive Reflection from Calculation.

Cognitive Reflection was also essentially uncorrelated with risk aversion and intertemporal choice in our experiment. This discrepancy from Frederick’s (2005) findings may be because choices in our experiment were incentivized, but CRT responses were not. One study has shown that incentivized predictors are more strongly related to incentivized outcomes, at least in the case of beliefs predicting behavior (Gächter & Renner, 2010). Without incentives, Cognitive Reflection may be a skill that helps avoid biases and errors in low stakes situations but may be less relevant in predicting incentivized choices because everybody reflects sufficiently. Numeracy is a requirement to resolve a mathematical problem in any situation, possibly explaining why it is a better predictor of these outcomes across levels of incentives.
General Discussion

Results of the present studies were consistent with the CRT’s role in decision-making biases and financial outcomes being due to numeric ability (Calculation) and not Cognitive Reflection. In addition, Study 1’s model of CRT responses indicated that Calculation accounted for much more of the variance in responses than Cognitive Reflection did. These results are at odds with previous explanations invoking the importance of intuitions and labeled the Cognitive Reflection Hypothesis in the present paper (Frederick, 2005; Toplak et al., 2011; Kahneman, 2003; Kahneman, 2013). The CRT either is not an effective measure of the hypothesized ability to check and correct intuitions or this ability does not play a role independent of numeric ability in the decision biases we examined. It is possible that Cognitive Reflection does play a role in other biases, such as probability matching, shown to be related to the CRT in previous literature. The present results, however, support the Numeracy Hypothesis, which posits that individuals with greater numeric ability will demonstrate fewer decision biases and achieve better financial outcomes, and it will account for the predictive power of Cognitive Reflection.

The three-item CRT scale remains a quick-to-administer predictor of a number of decision-making biases. It is also interesting psychologically. Analyzing the cognitive reflection aspect of this scale continues to lead to insights about human reasoning almost 10 years after publication of the initial paper (e.g., De Neys et al., 2013; Mastrogiorgio & Petracca, 2014). In addition, the fact that there are detectable individual differences in
Cognitive Reflection, that are somewhat stable across problems, may support the idea that individual differences in System 1 inhibition exist (Frederick, 2005; Kahneman & Frederick, 2002). These individual differences also may be related to executive inhibition, which itself relates to decision making in the lab (e.g., Del Missier et al., 2012) and in real life (e.g., Nigg et al., 2006; Roca et al., 2008). Theoretically, however, executive inhibition is distinct from Cognitive Reflection. The former measures the ability to inhibit a response, once it is clear that a response must be inhibited. The latter measures the ability to realize that a response should be inhibited in the first place (Toplak et al., 2011). Imagine two people choosing between a risky and a more favorable uncertain prospect (Ellsberg, 1961). The first chooses the risky option simply because he doesn’t like the feeling of uncertainty. The second reflects that the uncertain prospect is objectively a better choice, but chooses the risky option nonetheless. The first person is likely low on Cognitive Reflection, whereas the second is likely low on inhibition, but not on Cognitive Reflection. They are also distinct empirically; in particular, Toplak et al. (2011) found that CRT scores explain more of the variance in heuristics and biases than does inhibition (they did not control for numeric ability).

One possible reason for Cognitive Reflection not being a potent predictor is its lower reliability compared to Calculation. The latent variable associated with Cognitive Reflection accounted for just over half the variance as the latent variable associated with Calculation, indicating that inhibition of a default response on one item was not strongly related to inhibition of a default response on another item (i.e., not very reliable). This is
problematic for the CRT, because sum scores, which are often used as the outcome measure for the CRT, likely measure the more reliable Calculation construct to a larger degree than Cognitive Reflection. It also indicates that scoring schemes that differentiate only between intuitive and non-intuitive responses may ignore much of the useful variance (see also Pennycook, Cheyne, Koehler, & Fugelsang, 2015). However, this fact does not completely account for our results, since Cognitive Reflection has simple correlations with many of our dependent measures that then are accounted for by other factors (especially numeric ability) and its reliability was not much lower than that of numeracy (their respective Cronbach’s alphas were .54 and .65 in Study 1).

Based on the present results, correlations with CRT scores appear insufficient for establishing a prominent role for checking intuitions, at least in the decision tasks we examined and contrary to prominent citations of such correlations in support of this role (Sunstein & Thaler, 2005; Kahneman, 2013). Instead, the CRT scale appears to measure multiple constructs. At least two approaches exist to resolving the issue of multiple constructs: (1) Separate the hypothesized components of the scale mathematically or (2) use scales that measure only one construct at a time. The first approach was taken in the present two studies; it requires careful analysis of inattentive participants because component scores can be muddied by inattentive participants. In each of our studies, removal of these few participants (about 1%) did not significantly influence our results. Lower quality convenience samples often suffer from large proportions of inattentive participants, however (e.g., Maniaci & Rogge, 2014; Oppenheimer et al., 2009), and such
participants would be especially harmful to the Cognitive Reflection subscale. Techniques like robust regression can be used to automatically deal with such participants.

However, to attain a more pure measure of Cognitive Reflection, the second approach may be best: Scales that do not require the use of numeric skills should be used. Such scales would likely be less correlated with numeracy, and it would be interesting to see if they were uncorrelated when accounting for general intelligence. These scales could use problems that elicit an initial incorrect intuition but are not mathematical. Baron et al. (2014), for example, used syllogisms for this purpose.

The problem of multiple constructs may also apply to numeracy scales that include the original three CRT items (e.g., Weller et al., 2013). Performance on decision-making tasks may correlate with this numeracy scale due to Cognitive Reflection. However, the present results do not support this idea. In addition, four of five published studies of numeracy and CRT items supported them being part of a single numeric ability using factor analytic techniques (Låg et al., 2014; Baron et al., 2014; Weller et al., 2013; Liberali et al., 2012, Study 1).

Another issue exists, however, both for the CRT and for numeracy scales that include CRT items. The original three CRT items have been well publicized; they are commonly administered in internet surveys, have shown up in newspaper articles and radio shows, and are shown to undergraduates in courses. Problematically, experience with these problems is known to increase later performance (Chandler, Mueller, & Paolacci, 2014),
and it may test memory rather than performance. Indeed, studies recently conducted online show unusually high performance on the CRT (e.g., Mastrogiorgio & Petracca, 2014). In the present samples, it was unlikely that participants had prior exposure to the CRT items because Study 1 used mostly new CRT items (Toplak et al., 2014) and Study 2 took place in 2006, not long after the CRT was first published. In future studies, new CRT problems can be used, but it is currently unknown whether similar practice effects may exist with these new problems. In addition, because non-CRT numeracy problems were more potent predictors of decision biases and financial outcomes anyway, the best approach for future research may be to use these non-CRT numeracy items and systematically vary them while retaining similar difficulty levels.

One surprising finding was that, in both studies, numeracy did not fully account for the effects of Calculation in predicting decision-making biases. It may be that CRT Calculation indexes an aspect of numeracy, like algebraic ability, that was not otherwise measured in the Weller et al. (2013) numeracy scale. Alternatively, it may simply be, as Weller et al. found, that the remaining numeracy items were easier than CRT Calculation and that the added difficulty teases apart additional variance in decision biases among the most numerically able. Numeracy scales that separate various aspects of numeric ability may be useful (Ghazal, Cokely & Garcia-Retamero, 2014; see Weller et al., 2013 for discussion).

The proportions of variance explained in our studies were low to moderate (e.g., \( R^2 = 0.07 \) for the final model predicting Study 2’s financial outcomes). These results may be due, at
least in part, to the low reliabilities of the composite measures we used as dependent variables (though our reliabilities were not much lower than those of studies with similarly broad decision-making bias composites, e.g., Toplak et al., 2011). In addition, however, our composite measures represented multiple constructs. Individual items varied in how well they were predicted (see the simple correlations of Table 9). Although explaining more variance and having more reliable scales would indeed be desirable, numeracy nonetheless was an important predictor of decision biases and financial outcomes even after controlling for other variables (e.g., cognitive reflection, income). For example, although most people experienced primarily good financial outcomes (Mean=82% of good financial outcomes, Median=80%), a person who correctly answered one out of the 6 possible numeracy items was predicted to have 78.5% positive outcomes, whereas a person who scored 5 out of 6 correct was predicted to have 85.0% positive outcomes; this difference was enough to move a person from the 25th to the 65th percentile of financial outcomes in our sample. Given the crucial role these outcomes can play in life, the difference may be important. More important to the focus of the present paper, the data allow us to examine composites and biases for which Cognitive Reflection is a statistically robust predictor in simple correlations, but lacks any significant predictive power in the presence of numeracy.

Substantial research exists indicating that the CRT correlates with decision-making biases; various authors claim either cognitive reflection or general intelligence as explanations. The present results point instead to numeracy as a more important
explanatory construct. Future research should examine what specific aspects of numeracy matter to what kinds of decision making. For example, is knowing simple arithmetic sufficient? Peters et al. (2010), for example, used simple arithmetic problems to assess numeracy among Ghanaian villagers who did not know what abstract probabilities were; greater numeracy was associated with better decision-task performance and taking more protective health behaviors against HIV. Perhaps one must also know certain mathematical strategies and definitions for better decision making competence in some domains.

Peters and Bjalkebring (2015) suggested instead there are fundamentally different ways of knowing numbers. In particular, judgments and decisions are multiply determined by objective numeracy (associated with explicit number operations such as number comparisons and calculation), subjective numeracy (linked with motivation and confidence with the use of numbers), and number intuitions (the mapping of symbolic numbers onto magnitude representations was associated with numeric memory and valuation processes). Subjective numeracy (Fagerlin et al., 2007), beliefs in one’s own mathematical competence, may be a particularly overlooked measure. Being willing to work with numbers in decisions may distinguish between competent and incompetent decision makers more than being able to do the math (Peters & Bjalkebring, 2015). CRT research has been conducted primarily in the lab and is subject, therefore, to concerns about external validity. The same is true (although less so) for numeracy. The present research demonstrates, however, that CRT Calculation, as well as numeracy,
predicts decision-making competence in the lab and in real world outcomes for diverse samples of people.
Transition to Chapter 3

Chapter 2 introduced IRTrees, an analytic tool novel in the study of dual-system theories. This method allows CRT scores to be parsed into two sources of variance: differences in Cognitive Reflection and differences in Calculation. Given Calculation as a potential source of variance, past correlations with the CRT in the literature do not necessarily indicate that Cognitive Reflection plays a role in any of the studied biases. Instead, in the two studies of Chapter 2, we found that it is Calculation (and numeracy), not Cognitive Reflection, that is important. This finding, thus, casts doubt on previously reported correlations with the CRT as a source of empirical evidence for some dual-system explanations of decision-making biases. This evidence is especially problematic for default-interventionist theories which explicitly indicate that the ability to override intuitions is necessary to avoid biased decision making (e.g., Kahneman, 2003; Kahneman & Frederick 2002). It is, however, also problematic for any other theory that claims that biased decision making happens when associative processing overpowers deliberative processing (e.g., Epstein, 1994) if Cognitive Reflection is seen as a valid measure of the ability to prevent such an outcome.

It is less clear what the positive finding of the paper (that greater numeracy does predict less biased decision making) means for dual-system theories. On the one hand, numeracy, like Cognitive Reflection, could be seen as a measure of system 2 processing in some dual-system theories, so it could be seen as evidence for these dual-system theories. On the other hand, the numeracy scale may be predictive for multiple reasons. Math anxiety
and an approximate number sense are non-deliberative correlates (or perhaps components) of this competency that may have been responsible for its correlations with the decision biases and outcomes (Peters & Bjalkebring, 2015).

The lack of an effect of Cognitive Reflection on decision making when accounting for numeracy suggests that a different approach may be optimal for testing effects predicted by dual-system theories. In other words, dual-system theories may be accurate, but no effects of Cognitive Reflection were found in Chapter 2 because the CRT is a poor measure of the ability to inhibit intuitions. In addition, Chapter 2 serves as a reminder that correlational evidence always leaves open the possibility that the effect is due to a confounding variable (in the case of Cognitive Reflection, it appears to be numeracy). An experimental approach was chosen in Chapter 3 to avoid this possibility.

Although the review in Chapter 1 illustrated the variety of dual-system theories that exist, all of the reviewed default-interventionist and parallel-competitive dual-system theories converge on the point that processing in decisions under working-memory load should produce decisions more aligned with system 1. Manipulations of working-memory load (like increased time pressure or a concurrent task), therefore, are appropriate for testing these theories, and needed to draw conclusions about many of these theories simultaneously. However, the behavioral predictions of these dual-system models under load are unclear, even in common decision-making tasks, such as risky choice. For example, imagine a choice between a 30% chance to win $5 and a 40% chance to win $4. None of these non-quantitative models make it clear whether the former or latter option
will be favored to a greater degree when a working-memory load is present. One suggestion to deal with this dearth of specific predictions is to examine mathematical models that can make specific predictions in these tasks (Frank et al., 2009; Gawronski, Sherman, & Trope, 2014).

In Chapter 3, Mukherjee’s dual-system model (MDSM), a mathematical model of a parallel-competitive nature, is reviewed. Then, a study is conducted to test the possibility that working-memory load influences risky choices in a way consistent with this model’s predictions. The intention of the study is to elucidate one simple version of a parallel-competitive model and determine whether it is consistent with empirical data. An experiment guided by a mathematical model is used to address two common criticisms of dual-system models: that they are not clear or specific enough and that they do not make empirical predictions.

In terms of its mechanics, MDSM is most similar to Sloman’s (1996) parallel-competitive model, which, like MDSM, does not appear to allow interactions between the two systems in the model. It is less similar to Epstein’s (1994) model because that model allows complex interactions between the two systems (e.g., emotions influence deliberative thoughts and associative thoughts influence emotions). MDSM does not resemble default-interventionist models, in which system 1 processing must occur before system 2 is activated. MDSM is even less similar to more elaborately interactive models, in which the interaction is as important as or more important than the competition between systems (Loewenstein et al., 2001; Slovic et al., 2002).
CHAPTER 3: EXPERIMENTAL TEST OF DUAL-SYSTEM PREDICTIONS IN RISKY CHOICE

In the present chapter, I focus on the influence of a manipulation of working memory on the processing of information in risky choices through a cognitive modeling approach from Mukherjee (2010). As discussed in Chapter 1, the necessity of working memory for system 2 is a defining property of most dual-system theories (Evans & Stanovich, 2013). Specifically, many of these theories suggest that any influence of system 2 on processing should decrease or cease altogether as load increases. I focus on risky choice because it is historically one of the most prominent experimental paradigms used in tests of theories of decision making (Edwards, 1955; Kahneman & Tversky, 1979; Birnbaum, 2008).

However, few studies have been designed to directly test dual-system predictions in risky choice. In this chapter, I examine a mathematical model of a parallel-competitive theory of risky choice that does make such predictions (Mukherjee, 2010), along with modifications of the model. A study that tested these predictions using a working-memory load manipulation is presented and the model fits to the data are described.

Introduction

As summarized in Table 2, two of the major criticisms of dual-system models are that they are not clear or specific enough with respect to their structure and they do not make
specific predictions about behavior in a given experiment. Several researchers have suggested that one possible solution to these problems is a mathematical formulation of dual-system models (Frank, Cohen, & Sanfey, 2009; Gawronski, Sherman, & Trope, 2014). Most dual-system models are verbal; that is, they are specified using words, rather than equations. Although the study of decision making has often employed quantitative models, the advent of mathematical dual-system models is relatively recent. One dual-system model that is quantitative and therefore may have the most clear implications for decision research is Mukherjee’s Dual System Model (MDSM; 2010). Unfortunately, the model has never been fit to data and therefore its assumptions have gone untested. Trueblood (2013) extended this model to accommodate response times, allowing subsequent efforts to incorporate response time modeling, but she also did not fit her revised model to data. The advantage of studying such a mathematical model, compared to a verbal model, is that it makes very clear predictions that can be falsified. The disadvantage is that it is relatively new, and therefore not universally accepted by dual-system theorists who employ verbal models. As a result, any conclusions regarding this model may be specific to it. In the following sections, this model is explained, an addition to make it appropriate for fitting choice data is made (in the form of a logistic link function; Thurstone, 1927; Rieskamp, 2008), predictions are tested, and finally, the model is revised to more accurately fit the data.

MDSM was chosen because it is a quantitative dual-system model that can be applied to behavioral data. Mukherjee (2010) showed that, in addition to making its own
predictions, this model is able to account for a range of decision making biases at the group level, some of which are discussed below. It is the only quantitative dual-system model of which I am aware. At the same time, it is relatively simple computationally. As a result, it is uniquely positioned to help test dual-system model predictions for risky choices. As discussed, this is a specific parallel-competitive dual-system model. As a result, the validations and falsifications of its predictions may say little about other, less specific, parallel-competitive models, and they say even less about other types of dual-system models, like default-interventionist and interactive models.

Mukherjee’s (2010) Dual-system Model

Prospect Theory suggests that decision making in risky choice is best accounted for by a combination of a curved utility function and a curved probability weighting function (Kahneman & Tversky, 1979). MDSM can account for similar behavior, but is based in addition on previous research by Rottenstreich and Hsee (2001) and Hsee and Rottenstreich (2004). In their studies, evidence from multiple experiments indicated that the curvatures of these functions are more pronounced in affect-rich conditions (i.e., when the outcomes produce an emotional response) compared to affect-poor conditions (when the outcomes did not produce as much of an emotional response). Mukherjee’s model is parallel-competitive in that the values of the two systems are calculated separately and they are combined to create the total value of the gamble (see Figure 3). The value of system 1 (called the affective system) neither influences the value of system
2 (called the deliberative system), nor the other way around. In other words, this model is consistent with Figure 1’s simple parallel-competitive model.

According to MDSM, the greater curvature in affect-rich conditions arises from adding system 1 output to that of system 2. System 1 has a completely flat weighting function, weighting all non-zero probabilities equally, whereas system 2 has a normative, linear probability weighting function. Combining the outputs of these two systems produces a function akin to the Prospect Theory probability weighting function observed under high affect conditions. MDSM similarly suggests that the utility function is curved in system 1, but is linear in system 2. The model as suggested has two free parameters: the exponent of the value function, and the weight of the affective value of the gamble. This latter weight governs reliance on the system 1 and therefore apparent probability weighting. It is key to tests of MDSM: with greater weight of the affective value of the
gamble, one would expect probabilities to be ignored and amounts to be valued according to the exponent of the value function; this exponent is not expected to change with greater weight to the affective vs. deliberative value of the gamble. MDSM also has a constant, explained below. In MDSM, values of gambles are assessed as weighted sums of their affective and deliberative values.

$$V(G) = \gamma V_A(G) + (1 - \gamma)V_D(G)$$

where $V(G)$ is the value of some gamble $G$, and $\gamma$ is the weight of the affective value of the gamble, $V_A$ and $V_D$ are defined below. The affective value ($V_A$) of a gamble is calculated by exponentiating all possible outcomes of that gamble (that happen with non-zero probability) with the value function exponent and averaging these values. Since the values are averaged and probabilities ($0 < p < 1$) are ignored, $V_A$ is similar to a gamble value calculated using Prospect Theory, but with an extreme probability weighting function (e.g., Prelec’s, 1998, probability weighting function with parameter 0).

$$V_A = \frac{1}{n} \sum_{i=1}^{n} x_i^\alpha$$

where $x_i$ is the $i^{th}$ outcome of the gamble, $n$ is the number of outcomes in the gamble, and $\alpha$ is the exponent of the value function. For example, a certain win of $5 would be valued $\frac{1}{n} \sum_{i=1}^{1} 5^\alpha = 5^\alpha$.

The deliberative value ($V_D$) of a gamble is simply the expected utility multiplied by a constant, where $k$ is the constant and $p_i$ is the probability of the $i^{th}$ outcome of the
gamble. More detail on how the constant was handled is in the section about identifiability below.

\[ V_D = \sum_{i=1}^{n} p_i k x_i = k \sum_{i=1}^{n} p_i x_i \]

Essentially, the MDSM formulation is different from Prospect Theory in that subjective probability arises from a weighted sum of a normative, linear function (multiplied by the dollars to win) and a completely flat function that ignores differences in probability (multiplied by the subjective value of the dollars to win), rather than possessing an explicit functional form. This model can handle both gains and losses with only two free parameters. In comparison, Prospect Theory uses at least three parameters to handle gains and losses. In addition to making the model easier computationally, having fewer parameters restricts the model to make specific predictions discussed below.

The model, as shown, maintains monotonicity and transitivity, as well as transparent stochastic dominance. But, at the same time, it can account for a number of previously identified empirical inconsistencies, for example, violations of non-transparent stochastic dominance. This type of violation usually occurs when probabilities of outcomes are combined. For example, Birnbaum (1997, 2005) studies gambles like Gamble 1 (.90, $96; .05, $14; .05, $12) meaning 90% chance to win $96, 5% chance to win $14 and 5% chance to win $12, and Gamble 2 (.85, $96; .05, $90; .10, $12). Most people choose Gamble 2 because it has two “good” outcomes and one “bad” outcome, whereas Gamble 1 has only one good outcome. Gamble 1 stochastically dominates Gamble 2, however,
which can be made transparent by writing it (.85, $96; .05, $96; .05, $14; .05, $12) and writing Gamble 2 (.85, $96; .05, $90; .05, $12; .05, $12). Mukherjee’s model captures this inconsistency because, in its affective component, the probabilities are ignored and the outcomes are better in Gamble 2. Given that the weight of this affective component is greater than zero ($\gamma > 0$), one can assume that probabilities are attended to less than they should be, thus potentially capturing the inconsistency. Mukherjee shows that MDSM handles the full range of non-transparently presented stochastically dominated gambles. The model can also account for risk attitudes conforming to the Allais Paradox (Allais, 1979), the fourfold pattern of risk preference (Kahneman & Tversky, 1979) and other inconsistencies displayed in risk preference in the domains of gains and losses.

Different from most models of risky choice (e.g., Prospect Theory), this model can also account for ambiguity aversion and it makes new predictions regarding choices in which at least one option has an ambiguous component. To account for ambiguity aversion and several other relatively complex risky choice phenomena, the model makes it necessary to assume that the weight of the affective value of the gamble ($\gamma$) differs between two gamble options (effectively adding a third parameter). This assumption therefore makes it necessary to identify a priori which factors in a particular experiment can change $\gamma$ because post hoc decisions on this question can make the model undesirably flexible and introduce new free parameters. All wins in the present experiment will be relatively small (<$10). These winning outcomes were expected not to differ substantially in positive affect given prior research using similar ranges of outcomes when studying affect in
gambles and therefore it was reasonable to assume a constant $\gamma$ across outcomes (Bateman, Dent, Peters, Slovic, & Starmer, 2007); as reviewed below, however, $\gamma$ is expected to be different with vs. without working-memory load.

**Adding a link function.**

As proposed, the model does not explicitly include a choice mechanism. A deterministic choice strategy (that a person will choose the gamble with higher value) can be assumed. However, this is known to be a poor assumption: People choose a given option from a set with some likelihood, rather than deterministically (Mosteller & Nogee, 1951). Various functions can provide a mapping from values to probabilistic choices between options, to probabilistic choices of liking or not liking one option, to probabilistic continuous scales of judgment, etc. Since risky choice experiments often have participants make choices between two options, one appropriate link function providing the probability of choosing Gamble 1 over Gamble 2 is the inverse logit, which has seen wide use in the context of choice modeling from Thurstone (1927) to Rieskamp (2008). The function takes the form:

\[
P(G_1, G_2) = \frac{1}{1 + \exp(V(G_2) - V(G_1))}
\]

The addition of a scale parameter, $\delta$, can improve fit (Rieskamp, 2008). A scale parameter allows variability in the degree to which the value difference between gambles (as opposed to external factors) influences choice, thus accounting for imprecision in the choice mechanism due to factors like guessing. The following function (which is simply
the inverse logit with a scale parameter) can provide choice probabilities given values
provided by MDSM and will be used as the choice rule to fit to data:

\[ P(G_1, G_2) = \frac{1}{1 + \exp(\delta(V(G_2) - V(G_1)))} \]

In this particular experiment, it is especially important to include a scale parameter
because a potential difference between conditions is the frequency of guessing.
Participants may be more likely to guess due to being overwhelmed in the working-
memory-load condition than in the control condition. Since this difference is not value-
based, it should not influence parameters \( \alpha \) or \( \gamma \) and therefore should be explicitly
modeled. Nonetheless, in the model fit section below, I briefly explain a version of
MDSM without this additional parameter.

**Model identifiability with \( \delta \) and \( k \).**

Mukherjee provided little detail about either the psychological meaning of the constant \( k \)
(used in the modeled deliberative value of the gamble) or how it might influence model
fits. However, when \( 0 < k < 1 \), it might be used to balance the averages of
exponentiated values in the affective system and linear values in the deliberative system.
If \( k = 1 \), the affective system would undervalue its prospects compared to the
deliberative system because the parameter \( \alpha \), which is the exponent for the value is
typically less than 1. But if \( 0 < k < 1 \), system 1 will overweight small outcomes and
underweight large ones, consistent with the portrayal of system 1 in the diagrams in
Mukherjee’s (2010) manuscript. This parameter appears to be a convenience to improve
how meaningful other parameters are. Mukherjee referred to this parameter as a constant implying that its value should be set, rather than fit to data.

In the present instantiation of the model, it is especially important to set the value of this constant, rather than allowing it to be freely estimated because the parameter $\delta$ already performs a scaling function in the model (i.e., $\delta$ scales the difference in total value between two gambles, whereas $k$ scales the deliberative value), but $\delta$ is also more readily psychologically interpretable. If both parameters are allowed to vary in the estimation process, they will simply trade off and the model will fail to converge. For the purposes of the following experiment, $k = 3^{\alpha - 1}$, was used so that the deliberative and normative value functions would intersect at $3$, which was the median dollar value shown the participants in the experiment. This value would correspond to participants overweighting small winnings (<$3) and underweighting large ones (> $3), which is a desirable feature of the model, as it is a feature of diagrams in Mukherjee (2010).

Simulations confirmed that the model had trouble recovering parameters when $k$ was left as a free parameter. Parameter recovery was more successful and the model approximated the performance of Prospect Theory when $k$ was fixed to $k = 3^{\alpha - 1}$, even with $\delta$ in the model. The simulations also showed that when $\gamma$ was low (<.2), it was difficult to estimate $\alpha$, which makes sense because $\alpha$ does not influence choice when $\gamma$ is low (see Appendix B for more information on the simulations and an explanation of why the model was not identifiable when $k$ was left as a free parameter).
Manipulation of $\gamma$.

Although Mukherjee proposed that his parameter $\gamma$ should vary with increased or decreased affective content of the scenarios, this model can be used to assess any manipulation that is expected to increase or decrease reliance on either system. In MDSM, the weights of affect and deliberation sum to one, so more deliberation (higher $1 - \gamma$) necessarily means less affect (lower $\gamma$), and less deliberation means more affect. This is not an additional assumption but, rather, is one of MDSM’s features that allows it to have so few parameters. An alternative formulation of the model could have a second proportion parameter that would introduce other processes, such as heuristic evaluations, or responding based on the previous response, into the value function, in addition to the value-based affective system and the value-based deliberative system. In this alternative formulation, less deliberation would not necessarily result in more affective responding. Given the current formulation, however, MDSM makes several predictions about how a concurrent working-memory load should influence behavior in risky choice experiments.

Behavioral Predictions

Because a working-memory-load manipulation should be associated with greater reliance on the affective system based on dual-system theories, MDSM makes some qualitative behavioral predictions in this experiment. The choices used in the experiment are either between two gambles, each with two possible outcomes ($p$ chance to win $x$, otherwise nothing), or between one such gamble and another “gamble” with a certain outcome ($1$).
MDSM makes three strong behavioral predictions. These predictions are first stated and explained intuitively, then the mathematical relation of these predictions to the model is explicated.

Consider choices between $1 and a $p$ chance to win $x$. The former option, $1, is left alone by MDSM’s affective system ($1^\alpha = 1 \forall \alpha$). For the latter option, however, $x$ is exponentiated and averaged with $0$ in the affective system, which is the same as dividing the exponentiated value by $2$; it is also the same as replacing $p$ with a weight of .5. Because this happens regardless of $p$, it is a way of modeling the participant ignoring the likelihood of the gamble completely. So the affective system will underweight $p$ if $p > .5$, or keep it the same if $p = .5$. The affective system will also undervalue the outcome $x$. As a result, MDSM’s first prediction is that

Hypothesis 1 (H1). In the working-memory-load condition, the value of the gamble with $p > .5$ will be reduced relative to the value of the certain option and therefore the certain option will be chosen more often compared to the control condition.

Because probabilities are replaced by a weight of .5 in the affective system, $p$ will be overweighted in this system when there is a relatively small possibility to win ($p < .5$). For the same reasons as above, the outcome $x$ will still be undervalued in this system. The effects of overweighting $p$ and undervaluing $x$ will work against each other, and the difference in likelihoods of choosing the gamble between the two conditions will depend
on other parameters and on the values of $x$ and $p$. Therefore, the model’s second prediction is that:

Hypothesis 2 (H2): When $p < .5$, the difference between working-memory-load condition and control conditions in preference for the certain option will be smaller, compared to when when $p \geq .5$ or will be reversed. In other words, there will be an interaction between probability of winning the gamble and condition.

Finally, in MDSM, only one parameter, $\gamma$, governs reliance on system 1 vs. system 2 and therefore sensitivity to both likelihood and outcome amount. MDSM is designed to fit data in which sensitivity to probability is greatly reduced compared to expected value and sensitivity to amount is somewhere between unaffected (if $\alpha=1$) and reduced to the same degree as sensitivity to likelihood (if $\alpha=0$). If the manipulation decreases sensitivity to only the outcome and not probability, MDSM would have trouble fitting the data (because to reduce sensitivity to outcome through $\gamma$, the model has to reduce sensitivity to probability as well). Indeed, MDSM should have trouble fitting any data in which people are sensitive to likelihood but not amount, because neither of its components is thought to perform this function. Thus, this prediction provides a falsifiable test of the model. Thus, the model’s third prediction is that:

Hypothesis 3 (H3): Sensitivity to probability will be reduced to the same degree as or to a greater degree than sensitivity to amount in the working-memory-load condition vs. control.
These predictions can also be derived from the model mathematically. Consider the difference in value between the two gambles (the probability of choosing gamble 2 is a monotonically increasing function of this expression).

\[ V(G_2) - V(G_1) = \gamma V_A(G_2) + (1 - \gamma)V_D(G_2) - \gamma V_A(G_1) - (1 - \gamma)V_D(G_1) \]

Taking a derivative of this expression with respect to \( \gamma \) allows us to see how this value would change as \( \gamma \) increases. If this derivative is positive, we expect gamble 2 to be chosen more when a load is present; if it is negative, we expect gamble 1 to be chosen more when a load is present.

\[
\frac{d(V(G_2) - V(G_1))}{d\gamma} = V_A(G_2) - V_D(G_2) - V_A(G_1) + V_D(G_1) \\
= \frac{1}{n} \sum_{i=1}^{n} x_{2,i}^{\alpha} - k \sum_{i=1}^{n} p_{2,i} x_{2,i} - \frac{1}{n} \sum_{i=1}^{n} x_{1,i}^{\alpha} + k \sum_{i=1}^{n} p_{2,i} x_{2,i}
\]

When gamble 1 is a \( p \) chance to win $x$, and gamble 2 is $1$, the expression simplifies to

\[
\frac{d(V(G_2) - V(G_1))}{d\gamma} = 1 - k - \frac{x^{\alpha}}{2} + kpx
\]

This expression gets smaller as alpha gets larger. Let’s assume \( \alpha = 1 \) (the largest reasonable value). Then the expression

\[ 1 - k - \frac{x^\alpha}{2} + kpx \]

has a minimum at

\[ -\frac{x}{2} + px \]
which is still larger than 0 as long as \( p > .5 \). This means that regardless of the other parameters, as \( \gamma \) increases, preference for a certain option over a gamble with \( p > .5 \) will also increase, validating the first hypothesis.

However, when \( p < .5 \), \( 1 - k - \frac{x^\alpha}{2} + kpx \) can be positive or negative, but it will always be smaller than when \( p > .5 \) because the expression is monotonically increasing with respect to \( p \), validating the second hypothesis.

Finally, to assess whether we expect a cognitive load manipulation to influence probability sensitivity or amount sensitivity more, let’s take the derivative of \( 1 - k - \frac{x^\alpha}{2} + kpx \) with respect to \( x \) and \( p \) (these will reflect the rate of change with respect to amount and probability, respectively):

\[
\frac{dV(G_2) - V(G_1)}{dydx} = \frac{(\alpha - 1)}{2}x^{\alpha-1} + kp
\]

and

\[
\frac{dV(G_2) - V(G_1)}{dydp} = kx.
\]

When \( \alpha > 0 \) the latter value is clearly larger than the former, because \( x > 1 \) and \( \alpha - 1 < 0 \). This suggests that sensitivity to probability depends more on \( \gamma \) than does sensitivity to amount. Because \( \gamma \) reduces both, the model predicts that a load will influence sensitivity to probability more than sensitivity to amount (unless \( \alpha = 0 \), in which case it influences both equally).
Risky-choice findings relevant to working-memory-load manipulations

A number of previous studies have explored the link between working memory and risky choice using variations in deliberation. For example, more intelligent individuals generally have greater working memory; time pressure is sometimes thought to have similar effects to working-memory load in terms of reducing deliberation (Hammond, 1980). Although none of these studies have directly tested dual-system models, their findings may be relevant to (i.e., consistent or inconsistent with) these models. Specifically, MDSM predicts that working-memory load will influence sensitivity to probability and, depending on the parameters, amount of money (not sensitivity to amount alone, as discussed). If findings of these time-pressure and intelligence studies are consistent with MDSM’s predictions, they may inform expectations of MDSM accuracy.

To begin, Figner et al. (2009) examined a gambling task in which participants received a reward for drawing cards from a single deck. The participants were told that the deck contained several cards that would result in a loss (in fact, there were no losing cards in the deck). Given this information, there existed an optimal, non-zero number of cards to flip in order to maximize expected winnings. This number depended on the number of cards participants were told would result in a loss, the amount to lose on the losing cards, and the amount to win on the winning cards. The authors calculated a complexity of information score for each participant, indicating how many of these three pieces of information significantly predicted that participant’s decisions. Less intelligent
participants (those with lower IQ scores) paid attention to fewer different pieces of information about the deck (the authors did not examine which pieces of information less intelligent participants ignored), and they were more likely to gamble. As a result, the less intelligent participants’ decision making was thought to be less optimal, consistent with dual-system models. This behavior is consistent with MDSM, which predicts that participants using system 2 to a lesser degree, as less intelligent participants may be expected to do, will ignore certain features of gambles. However, this experiment did not test MDSM’s more specific predictions regarding the degree to which probabilities vs. amounts are ignored.

Second, Young et al. (2012) examined the use of probability vs. outcome information under various levels of cognitive load. They studied certainty equivalents for simple gambles (i.e., they asked participants for the largest amount of money they would give up in order to play each gamble). The outcome of the gamble depended either on a random event or on the correctness of one’s answer to a trivia question. The authors analyzed the results in a Prospect Theory framework, fitting a value function and a two-parameter probability weighting function. These authors found that, in both the random event and trivia question conditions, increased time pressure resulted in less sensitivity to probability, but had no effect on the value function. This finding is consistent with MDSM, which predicts that when people rely on system 1, probabilities would be influenced to a greater degree than amounts would.
Third, Payne, Bettman, and Johnson (1988) examined the use of heuristic processing in decisions between complex four-outcome gambles in a process-tracing task. One of their findings was that increased time pressure decreased the proportion of time spent examining (i.e., attention to) outcomes slightly and therefore increased the proportion of time spent examining probabilities. They did not examine the use of probability vs. outcome information, however. Nonetheless, their attention results may be inconsistent with MDSM, which predicts that as deliberativeness decreases, likelihoods would be ignored at least as much as amounts.

In the present study, the methods were chosen first and foremost to provide a good test of MDSM’s predictions. Instead of time pressure or individual differences, we employed a working-memory-load manipulation, but this manipulation is conceptually similar to the manipulations and individual differences described above.

**Method**

**Sample**

A sample of 94 undergraduates took part in this experiment. Data were collected throughout spring 2016. The experiment continued until the end of the semester, which was used as the stopping rule for data collection. The posting informed participants of the possibility that they would receive a small sum of money.

Although demographic information was not collected from these specific participants to save time for the lengthy main task, it was a sample from undergraduates in Ohio State
University who are 72% white, 53% male, 76% Ohioans, and have a median age of 20 years.

**Procedure**

Participants were randomly assigned to the working-memory-load or control condition and participated in the experiment one at a time. Instructions were read aloud to participants and they confirmed their understanding at several points. After agreeing to participate, all participants were given practice with a working-memory-load task in which they identified when consecutive tones were identical. There were four possible easily distinguishable tones. Tones were presented one at a time and the next tone was randomly selected, subject to the constraint that there were never 3 identical tones in a row. This constraint was imposed because responding to the third identical tone was confusing to participants in pretesting. If the tone currently presented was identical to the previous tone, the participant was told to hit the spacebar before the next tone sounded (there were 1300 ms between tones). Once the participant hit the spacebar after a tone repeated twice in a row, but before the next tone sounded 10 times without making any errors of omission or commission, this phase ended. This part of the experiment served as practice for the competing task for the participants in the working-memory-load condition, but the participants in the control condition also performed it to keep the experiment identical between conditions with the exception of the manipulation.
This working-memory task was chosen because pretests showed that it was able to decrease accuracy of responses in a second task where participants were explicitly asked to choose the option with a higher expected value. Since the Mukherjee model assumes an expected value calculation occurs in system 2, this task was an effective manipulation of reliance on this system.

Participants were next told that they would make a series of choices between gambles and that one choice would be selected at the end of the experiment and honored. Participants were given four practice trials, representing the range of gambles in the experiment. Participants then made choices between 100 pairs of monetary gambles (sometimes a “gamble” was a sure $1). One trial was randomly chosen at the end of the experiment, and the participant’s choice was honored (incentivizing one of many choices is a standard practice in incentivizing choices between gambles; Holt & Laury, 2002). For example, if a 30% chance to win $4 was chosen, a 10 sided die was rolled, and the participant was awarded $4 if the outcome of the roll was 1 2 or 3; otherwise, the participant was not paid (also subject to the constraint below). Participants were informed that this is how they would be paid and were given the 10-sided die to examine.

Half the participants were told they would have to perform successfully the earlier tones task while selecting the gambles. After two consecutive identical tones, these participants had to hit the spacebar before the next tone sounded (1300ms), like in the practice task above. When a trial was chosen at the end, these participants were not rewarded if they had missed a tone during that trial. This was an appropriate incentivization scheme
because participants had to pay attention to the competing task, while still making choices consistent with their preferences. Without incentivization, participants would be likely to ignore one of the two tasks in order to reduce cognitive load. Participants responded to the tones using the spacebar as in the practice trials, and they responded to the gambles, using the F and J keys, also as in the practice trials.

**Design**

Fifty of the total 100 choices were between gambles and a 100% chance to win $1 (e.g., 10% chance of winning $8 vs 100% chance of winning $1). The gambles were randomly selected from a pool of 81 (9x9) possible gambles with a probability to win [.1, .2, .3 … .8, .9] and amount to win [$1, $2, $3, … $8, $9]. However, the pool was further restricted, such that the absolute difference between the expected value of the gamble and $1, divided by the smaller of these two quantities was less than 3, leaving 59 gambles (a similar method was used by Rieskamp, 2008, to select gambles). This restriction gets rid of some really obvious choices (e.g., a choice between 90% to win $9 vs. 100% to win $1), but leaves in other obvious choices. Some easy choices (that had a big difference in expected value) and some dominated choices were left in so that people in the working memory condition were not incentivized to choose randomly (if all expected values were similar, the incentive may be to attend to the working memory task only). Choices of dominated options can also be used to gauge attention.
The remaining 50 choices were between pairs randomly selected from the pool of all possible 3,240 pairs of the 81 gambles ($\binom{81}{2} = 3,240$, the number of ways 81 gambles can be paired). This pool was further restricted, like above, such that the absolute difference between the expected values of the gambles, divided by the smaller expected value was less than 3. This left 2,189 pairs of gambles.

The same 100 randomly selected gambles were used for all participants. This choice was made to cut down on unnecessary variance in the experimental design. The order of presentation, however, and counterbalancing from left to right, were randomized for each subject in order to eliminate systematic order effects.

**Results**

In the control condition, 46/47 participants answered 1 or 0 dominated gamble trials incorrectly (out of 7), with the median participant answering all trials correctly (mean = 97%). The remaining participant answered only 17% of dominated gambles correctly and gave an inappropriate response in another choice with a dominated option. In the working-memory-load condition 29/47 participants answered 1 or 0 incorrectly, but the median participant answered 92% of the dominated gamble trials correctly (mean = 88%, range=[50%,100%]). In the working memory condition, participants more often missed a repeating tone (False Negative Rate=22%, range=[0%,98%]) than reported a repeating tone when one was absent (False Positive Rate=3%, range=[0%,15%]). Although the ranges indicate that a few participants ignored one task or the other, all participants were left in for the analysis for the following four reasons. First, I wanted to avoid setting
arbitrary cutoff criteria. Second, it is difficult to distinguish inattentive participants from those having difficulty performing the difficult task assigned in the experiment. Third, the hierarchical modeling approach used below is relatively robust to outliers, especially because a guessing parameter was included and Cauchy distributions were used for priors. Cauchy distributions have infinite variances and very thick tails, handling outliers effectively.

The results of the experiment for the non-dominated gambles are depicted in Figure 4, which presents the results for choices between a certain $1 and a gamble in its left panel, and the results for choices between two gambles in its right panel. Data from choices between a gamble and $1 for sure (on the left) are broken up by the amount that could be won in the gamble. The working-memory-load condition is in gray and the control condition is in black. The probability of winning in the gamble is on the x-axis and the proportion choosing the certain option is on the y-axis. The negative slopes of all lines indicate that, as the probability of winning the gamble increased, the proportion choosing the certain option decreased. Data from choices between two gambles (on the right) are broken up by the difference in amounts that could be won in the two gambles. The difference between probabilities of winning in the two options is on the x-axis and the proportion choosing the higher-likelihood gamble is on the y-axis. The positive slopes of all lines indicate that as the difference in probability of winning the gambles increased, the proportion of people choosing the gamble with the higher likelihood also increased.
Figure 4. Mean proportions and 95% confidence intervals by condition of those who chose the certain option over a gamble (left-side panels) and those who chose the higher likelihood gamble among two gambles (on the right). A higher proportion choosing the certain option (in the left panels) necessarily means a lower proportion choosing the gamble and a higher proportion choosing the higher likelihood gamble (in the right panels) means a lower proportion choosing the higher dollar outcome gamble.
These patterns of responding are partly consistent with MDSM’s three major behavioral predictions. The model’s first prediction (H1 that people would choose the certain option more often in the working memory condition and therefore, high-probability gambles less frequently) was supported (see the right sides in the left-hand column of Figure 4). For example, the 90% to win $4 gamble, a gamble with the maximum expected value difference from $1, was chosen over the $1 for sure by 89% of people in the control condition, but by only 67% of people in the working-memory-load condition (t=2.3; p =.02). In addition, H2 (that when p < .5, preference for the certain option may still be higher in the working memory condition, but the difference between conditions will be smaller, or the preference for the certain option will be lower in the working memory condition) was also validated, as indicated by the left side of Figure 4. In particular, the effect of condition visible on the right side disappears or reverses on the left side (preference for the certain option was lower in the working memory condition or about equal to the control). For example, the 10% to win $4 gamble was chosen over the $1 for sure by 9% in the control condition, but by 73% in the working-memory-load condition (p=.02), a reversal of the effect described above.

The model’s third prediction (H3) was that people should neglect likelihood more than money in the working memory condition relative to control. However, the data did not support this prediction. Instead, money was ignored to a greater degree than likelihood in the working memory condition. To illustrate this result, first consider choices between a 40% chance to win $4 and a 30% chance to win $6. In the control condition 28% of
people preferred the low outcome, high probability alternative (a 40% chance to win $4), compared to 63% in the working-memory-load condition (t=3.2, p<.01). This finding indicates that in the working-memory condition, people preferred a small advantage in likelihood over a small advantage in amount to a greater degree than in the control condition. Of course, an alternative explanation is that people had the same preferences, but simply guessed more in the working memory condition (63% was not significantly different from choosing at a random 50%; p=.10). But consider choices between a 90% chance to win $2 and a 10% chance to win $8. In the control condition, 83% of people preferred the former high money, low probability gamble. If people are choosing at random in the working memory condition, we would expect this percentage to decrease. Instead, there was no significant change: 80% of people preferred the former gamble in the working memory condition (t=.151, p=.88). The results suggest that people use probabilities, rather than monetary amounts, to a greater degree when a working-memory load is present. This effect can be observed in Figure 4 by noticing that working-memory load increased the proportion of high-likelihood gambles chosen on the left side of the right-hand panels of Figure 4 (similar to how it increased preference for a 40% chance to win $4 over a 30% chance to win $6) without reducing it on the right side of the same panels (similar to how it produced no change in preference for a 90% chance to win $2 over a 10% chance to win $8). The exception is the top panel in the right column, where there appears to be a slight decrease in the proportion of people who choose the higher likelihood gamble in the load condition, compared to control. However, there were no
gambles for which the difference between the two conditions was statistically significant in this direction.

People using likelihoods more with greater load is inconsistent with MDSM, which predicts that as deliberative capacity decreases, people should ignore likelihoods at least to the same degree that they ignore amounts. Instead, as the opportunity to deliberate decreased, people appeared to become more fixated on likelihoods, preferring gambles with higher likelihoods to a greater degree, even when the difference in likelihood was small. In order to assess the accuracy of the model’s predictions to all the gambles in the experiment (instead of just those described above), we fit the model’s parameters to data.

**MDSM model fits**

All models described were fit in a Bayesian hierarchical framework. Half of the trials in the experiment were randomly selected for model fit and half left out for validation. First, an attempt was made to fit the data without the guessing parameter, $\delta$, adding only the inverse logit link (to provide a mapping between values and probabilistic choices) to the model described in Mukherjee’s (2010) article. However, as described in Appendix C, this exercise was unsuccessful, since this version of the model could not fit many of the patterns in the data.

In a second attempt to fit the data, participants were each given their own set of parameters $\alpha$, $\gamma$, and $\delta$ (see Figure 5 for a graphical depiction of this model; see Appendix D for the code).
Figure 5. A graphical depiction of the Bayesian MDSM. Shaded boxes are observed discrete quantities, circles with solid outlines are continuous parameters and circles with double outlines are continuous derived quantities. Index $g$ is for gamble, $t$ is for trial, $s$ is for subject and $c$ is for condition.

**Parameter Distributions**

A Beta distribution was used for the subject-specific $\gamma$ parameter ($\gamma_{s,c}$), because this distribution is appropriate for parameters that represent proportions and are constrained...
between 0 and 1, both of which are properties of \( \gamma \). The mean of this parameter, \( \mu_{\gamma,c} \), was also modeled using a Beta distribution and depended on the condition. Its standard deviation, \( \sigma_{\gamma,c} \), had a Cauchy prior with a mean of 10 and a spread of 20 and was not assumed to be equal between conditions. Cauchy priors are a recommended choice for priors on variance parameters in hierarchical models (Gelman, 2006). However, Cauchy distributions have infinite support, so the standard deviation parameters had fixed minima at 0, resulting in truncated Cauchy priors. The grand mean of \( \gamma \), \( (\mu_{\gamma}) \) had a uniform prior between 0 and 1. The standard deviation of the mean \( \sigma_{\gamma} \) had a Cauchy prior with a mean of 3 and a spread of 20.

A gamma distribution was used for the subject-specific \( \alpha \) parameter \( (\alpha_{s,c}) \), because this distribution is appropriate for parameters that have to be greater than 0. The mean of this parameter \( (\mu_{\alpha}) \) had a Cauchy prior with a mean of \( \frac{1}{2} \) and a spread of 1, whereas its standard deviation \( (\sigma_{\alpha}) \) had a Cauchy prior with a mean of 1 and a standard deviation of 2.

A gamma distribution was also used for the subject-specific \( \delta \) parameter \( (\delta_{s,c}) \). The mean of this parameter, \( \mu_{\delta,c} \), was modeled using a Cauchy distribution and depended on the condition. Its standard deviation, \( \sigma_{\delta,c} \), had a Cauchy prior with a mean of 50 and a spread of 150 and was not assumed to be equal between conditions. The grand mean of \( \delta \), \( (\mu_{\delta}) \) had a Cauchy prior with a mean of 4 and a spread of 20. The standard deviation of the mean \( \sigma_{\delta} \) had a Cauchy prior with a mean of 10 and a spread of 50.
Cauchy priors and distributions were used where possible because they have infinite variance and are very diffuse as a result. Diffuse priors were appropriate because this was the first time Mukherjee’s model was fit to data and because they attenuates the influence of outliers, which were likely present in this dataset. All parameters had fixed minima at 0 because the model is no longer psychologically plausible if means or standard deviations or standard errors of parameters are negative (negative $\alpha$ would mean preferring smaller amounts of money to larger ones in system 1 and negative $\delta$ would mean being more likely to choose options that you value less).

The model was fit in Stan (Stan Development Team, 2016) with step size reduced and tree depth increased in order to accommodate the correlated parameters (means and standard errors for some parameters were highly correlated inhibiting the sampler’s ability to make jumps with default settings). The model was stopped after 3,000 warm-up and 3,000 sampling iterations of 4 chains, yielding convergence on all relevant parameters, as assessed by the $\hat{r}$-statistic (Gelman & Shirley, 2011). The $\hat{r}$-statistic assesses the stability of outcomes between and within multiple chains. Values greater than 1.1 indicate lack of convergence. Values of $\hat{r}$ for all parameters in the present simulation were lower than 1.05.

**Model Results**

The posterior mean in the load condition of $\gamma$ was slightly higher $\mu_{\gamma_{\text{load}}} = .44$ than in the control condition $\mu_{\gamma_{\text{control}}} = .42$. Figure 6 shows the uncertainty in the posteriors of all
key parameters. The difference between conditions in $\gamma$ is very small compared to the uncertainty, but the difference in $\delta$ is quite large. One surprising outcome of the model fit was that the exponent of the value function, the $\alpha$ parameter, was near 0 for a majority of participants ($\text{mean}(\alpha_{s,c}) < .05$ for 60% of participants, $\text{mean}(\alpha_{s,c}) < .10$ for 75%), essentially reducing the value in the affective system of the model to a preference for certainty, regardless of the properties of the gamble, and to a preference that was nearly indifferent between any two gambles. This outcome occurred because the affective portion reduces to 1 for the sure $1 options and to $\frac{1}{2} d^\alpha \approx \frac{1}{2}$ for all gambles when $\alpha$ is sufficiently small.
Figure 6. Priors and posteriors on key MDSM parameters in each condition. Posteriors indicate differences in $\gamma$ are largely indistinguishable from uncertainty, whereas differences in $\delta$ are noticeable. Low $\alpha$ values indicate very low modeled sensitivity to amount.
After fitting MDSM, its predictions for the remaining 50 gambles for each participant were simulated. It correctly predicted 73% of choices with means of parameter posteriors (compared, to for example, always choosing the option with the higher likelihood, which predicted 53% of choices correctly, or basing choices on expected value, which predicted 65% of choices). As can be seen in Figures 7 and 8, MDSM was able to reproduce some differences between conditions in the test data. However, most of the differences between groups in model fits were due to the $\delta$ parameter for random choice, which, in the working-memory-load condition, was less than half its value in the control condition and indicated more guessing. Namely, both in the data and in the model, the lines in the working memory condition tend to be flatter than those in the control condition. However, the model systematically underestimated sensitivity to probability in both conditions. This is evident from the model estimates consistently underestimating the frequency with which high probability gambles were chosen (see the right side of Figure 6 and all of Figure 7). The intuition for the model’s inability to fit these data is that it has only one parameter, $\gamma$, governing sensitivity to both likelihood and amount. Therefore, with a non-zero $\alpha$, as $\gamma$ increases, the model predicts sensitivity to probability will fall more rapidly than sensitivity to amount. With a zero $\alpha$, the model predicts sensitivity to probability and amount will fall at the same rate. The model cannot account for greater sensitivity to likelihood than amount, which was observed in the data for most participants. The closest it can get is reducing $\alpha$ close to 0, which it did for most people.
Figure 7. Mean proportion choosing the certain option (from the validation data) and estimated proportions choosing the certain option (from MDSM model fits) based on choices between a gamble and $1 for sure. MDSM overestimated the proportion of people choosing the certain option (and therefore underestimated the proportion choosing the gamble) when the probability of winning was >.6.
Figure 8. Mean proportion choosing a higher likelihood gamble over a lower likelihood gamble (from the validation data) and estimated proportions choosing the higher likelihood gamble (from MDSM model fits) based on choices between two gambles. MDSM underestimated the proportion choosing the higher likelihood gamble.
**Alternative model fits**

In order to ascertain whether it is the inability to account for probability sensitivity that produces the model’s systematic deviations from data, an alternative version of MDSM was examined. This version assumes that the affective system weights the probability and instead ignores the amount of money:

$$V_A = \sum_{i|d_i \neq 0} w(p_i, \alpha)$$

This formulation of the affective system takes all possible non-zero outcomes in a gamble and weights their likelihoods using a function $w$ and then sums the weighted likelihoods. This change in the affective system also required a change in the constant, $k$, in the deliberative system. The reciprocal of the average outcome is appropriate for the constant, $k$, in the deliberative system because it would set the deliberative and affective values for the average outcome to be equal, as was done for the original MDSM. In this experiment, 3 was the median outcome, so $k = \frac{1}{3}$ was used. The deliberative system and the way the valuations of the two systems are combined are otherwise identical to MDSM. Prelec’s (1998) probability weighting function was used in the affective system because people might be expected to perceive probabilities non-linearly in the affective system instead of ignoring them completely, and this function has seen wide use as a probability weighting function:

$$w(p, \alpha) = \exp\left(-(-\log(p))^{\alpha}\right)$$
The alternative version of the model was fit in the same way as the original MDSM to the same training data (see Figure 9) with the changes to the affective value function and constant as described above. Like in the original MDSM, $\gamma$ was higher in the working-memory-load condition than in the control. However, this time, the difference in $\gamma$ between conditions was $.08$ ($\mu_{\gamma,load} = .42$; $\mu_{\gamma,control} = .34$), which is 4 times higher than the difference in $\gamma$ between conditions in the original MDSM. Another feature of the model fits was that the parameter $\alpha$ was around or above 1 for most participants, indicating that instead of the usual reverse-s-shaped probability weighting function which overweights small probabilities and underweights large ones, the function was relatively linear in the affective system, or even slightly s-shaped, slightly underweighting small probabilities and overweighting large ones. This function indicates that when using the affective system, instead of ignoring probabilities altogether, participants either weighted them appropriately or even over-relied on differences between small and large probabilities in the examined range (.1 - .9).
Figure 9. Priors and posteriors on key modified model parameters. Posteriors indicate noticeable differences in $\delta$ and $\gamma$ between conditions and that $\alpha$ was likely greater than 1, indicating high sensitivity to likelihoods.
In terms of fitting the validation data, the altered MDSM model performed well, it correctly predicted 77% of choices with means of parameter posteriors, only a small improvement over the original MDSM (73%). However, unlike MDSM, it largely reproduced the patterns in the data (see Figures 10 and 11). Mean posterior parameters were found for each participant using the original MDSM and the modified version. The using these parameters to predict each participant’s choices, log likelihood for the original MDSM was -2120; for the modified MDSM, it was -1911, while using the same number of parameters. This difference indicates very strong preference for the latter model according to a likelihood ratio test (p <.01; Fox, 1997). The one exception to the good fits of this version of MDSM was for gambles where the outcomes differed by a large amount. This is evident in the bottom panel of Figure 11, where the model slightly underestimated the frequency with which people chose the high-outcome gambles (and therefore overestimated the frequency which which they chose high-likelihood gambles). It appears that when the difference between the outcomes was very large, they were used to some degree even by the affective system.

10 Comparing Bayesian models by conducting a likelihood ratio test on posterior probabilities is one of many ways Bayesian models can be compared (Gelman, Meng & Stern, 1996; Gelman, Carlin, Stern, & Rubin, 2014). No one method for model comparison is best, but given the overwhelming strength of evidence using this test, it is unlikely that the type of test will matter.
Figure 10. Mean proportion choosing a gamble over a certain option (from the validation data) and estimated proportions choosing the certain option (from modified model fits) based on choices between a gamble and a certain option. The model that ignores outcomes instead of probabilities fit patterns in the data adequately.
Figure 11. Mean proportion choosing a higher likelihood gamble over a lower likelihood gamble (from the validation data) and estimated proportions choosing the higher likelihood gamble (from modified model fits) based on choices between two gambles. The model that ignores outcomes instead of probabilities fit patterns in the data adequately, except when the difference in the outcomes between gambles was very large. In those cases, it slightly overestimated the proportion choosing the higher likelihood gamble (and therefore, slightly underestimated the proportion choosing the higher dollar amount).
Is the $\delta$ parameter doing all the work?

One possible concern about this model is that its specific dual-system structure is irrelevant to fitting the data. Specifically, it is possible that the data can be effectively modeled by any value-based process with the addition of random noise in the $\delta$ parameter. To test this idea, a simple expected value plus noise model was fit (Rieskamp, 2008). Although this model did appear to fit some of the differences between conditions well, it performed poorly in fitting the actual data, severely underestimating the proportion of certain options and high-likelihood gambles that were chosen (see Appendix E).

Discussion

Chapter 1 introduced a number of criticisms raised by other researchers that concern dual-system models. Some of the most substantial criticisms highlighted that dual-system models do not make enough predictions about behavior and are not clear or specific enough about their structures. Chapter 2 attempted to test, but failed to find evidence for, dual-system predictions based on a default-interventionist perspective by examining the CRT, a popular scale. However, failing to find evidence for a dual-system model using the CRT does not necessarily indicate that such a dual-system model is wrong because these results could be due to a lack of power or because the CRT is a bad scale. In Chapter 3, a quantitative, albeit simplified, parallel-competitive model was fit to data. Again, this model made strong predictions, but some of these predictions were not consistent with the data. For example, the model predicted that, when unable to use the
deliberative system, people would ignore likelihoods, whereas the experiment showed that people continued to rely on them, but instead neglected outcomes. This evidence contradicts a part of the model, rather than simply failing to find evidence consistent with the model. However, it supports the general idea of greater reliance on a heuristic process under working-memory load, though the specific heuristic process proposed by Mukherjee (2010), as operationalized here, did not appear to apply in this case.

An alternative dual-system model that could account for these findings was proposed. This alternative model did fit the major patterns in the data from this experiment by relying on increased use of probabilities as well as increased guessing in the affective system. Unlike the original formulation, this version is consistent with empirical findings that suggest that even primates are sensitive to different levels of probability of a constant outcome (Fiorillo et al., 2003) and that evolutionarily older areas of the brain are responsible for at least some of this probability sensitivity in humans (Preuschoff et al., 2006). These findings suggest that any affective system should have at least some sensitivity to probability. It should be noted that a similar criticism can be applied to the present formulation (animals can clearly distinguish between amounts), but at least the present formulation is one that generally fits the empirical data. The model should be developed and tested further before being applied in a large set of circumstances.

In retrospect, this pattern of results is consistent with two lines of previous research. Payne, Bettman, and Johnson (1988) found that, as time pressure increased, decision makers tended to shift to simpler strategies and base decisions on attributes manipulated
to be normatively important, while neglecting other, less important attributes. In the present experiment, gambles were not chosen so that probabilities or outcomes would result in a normative advantage (if anything, expected value was more correlated with outcomes in the gambles that were selected). However, Slovic and Lichtenstein (1968) found that probabilities are the default primary attribute of a gamble as assessed in binary choice, both in the domain of gains and in the domain of losses (see also Slovic, 1967). This was true even when the gambles were presented pictorially instead of the p% to win $x format used in this study. Similarly, Bateman et al. (2007) had people rate gambles which varied in probability and in outcome on a subjective attractiveness scale. They found that these subjective ratings depended on probability to a much greater degree than on outcome although the expected value of the assessed gambles depended equally on the outcome and probability. Research on evaluability suggests that this may be because probabilities are easily evaluated due to their bounded range, whereas amounts of money are more difficult to evaluate (Hsee, 1996).

Thinking about these results in the light of people’s tendency to rely on important features to a greater degree when a working-memory load is present also explains the apparent discrepancy between these results and those observed by Young et al. (2012). Those authors found that, when under time pressure, participants’ probability weighting function changed whereas their value function did not. Consider that Slovic and Lichtenstein (1968; see also Lichtenstein & Slovic, 1971; Grether & Plott, 1979) found preference reversals between pricing and choice: When pricing gambles, people tended to
care about dollar amounts, but when choosing, they tended to care about probabilities. Young et al. (2012) examined a pricing task, whereas choices were examined in the present experiment. If people have a preference to use a certain attribute in a particular task type, working-memory load appears to increase that preference.

The proposed model is new and has several limitations. First, unlike the original MDSM, it is thus far unclear exactly which behavioral biases this model is theoretically able to account for. It should still be able to handle risk aversion for gains. For example, $0.95 (0.95 in the deliberative system and 1 in the intuitive system) will be preferred to a 95% chance to win $1 (0.95 in the deliberative system, and .95 in the intuitive system, assuming $\alpha=1$ and $k=1$, but other reasonable parameters would also work). It should also be able to handle risk seeking for losses. For example, a 95% chance to lose $1 (-0.95 in the deliberative system, and -.95 in the intuitive system, again assuming $\alpha=1$ and $k=1$, but not excluding other reasonable parameters) will be preferred to losing $0.95 for certain (-0.95 in the deliberative system and -1 in the intuitive system). However, accounting for common consequence and/or common ratio effects may be trickier and depend on the precise parameters.

Second, although the present experiment used a diverse set of gambles of two types, these gambles did not nearly cover the entire range of possible gambles and they were presented in a specific way. For example, gambles with smaller (<10%) or larger (>90%) probabilities than those studied in this experiment may be overweighted and underweighted, respectively, and thereby change the optimal parameters for the model or
require a different formulation of the probability weighting function to handle. In addition, though we presented the gambles in a standard, neutral way, other forms of presentation may yield other results. For example, highlighting the dollar amount may cause people to ignore the probability in the working memory condition. Another way the presented results may be flipped is if probability varies much less than the outcome in the experimental stimuli, and therefore contributes less to expected value. This kind of experiment would incentivize an optimal decision maker who can only concentrate on one attribute to ignore probabilities. Finally, this experiment did not examine framing effects, or losses, instead concentrating on the domain of gains.
CHAPTER 4: GENERAL DISCUSSION

The present research reviews dual-system approaches to decision making, as well as their limitations, and examines the utility of two analytic approaches for testing dual-system theories of decision making. It therefore attempts to address some of the stated limitations of dual-system models. In particular, Chapter 1 reviewed these dual-system approaches, organizing them into default-interventionist, parallel-competitive, and interactive types. The theories are often criticized for being vague and not making specific predictions. However, in Chapters 2 and 3, specific predictions of two of these models were tested. Chapter 2 used IRTrees (De Boeck & Partchev, 2012) as an analytic approach well-suited for testing default-interventionist models, one type of dual-system theory. IRTrees involve separating responses based on the processes that give rise to them, allowing measurement of the interventionist propensity for cognitive reflection more directly than other approaches. This method of analysis failed to provide evidence for default-interventionist models across a large range of decision-making biases and outcomes. Specifically, and contrary to default-interventionist models’ predictions as well as prior claims in the literature, individual differences in cognitive reflection did not independently predict individual differences in the examined biases and outcomes.
Chapter 3 examined Mukherjee’s Dual-System Model (MDSM), testing qualitative and quantitative predictions in a risky-choice experiment in which cognitive load was manipulated. This parallel-competitive model did not provide an accurate account of the data. Instead, it underestimated the degree to which people rely on probabilities in risky choice overall and under cognitive load in particular. A similar parallel-competitive model which did account for these patterns was proposed.

Evidence from Chapters 2 and 3 for and against Dual-System Theories
The findings in both chapters rarely provided support for the specific dual-system theory tested. Chapter 2 failed to provide evidence for any type of theory that suggests that inhibiting intuitions or emotions is vital to good decision making. By definition, default-interventionist theories posit such a mechanism and, in fact, some have used correlations with the CRT as supportive evidence for these theories (Kahneman, 2003; 2011; Stanovich, 1999; 2004; Evans, 1984; 1989; 2006). Chapter 2’s data suggests that this use is inappropriate. By contrast, simple parallel-competitive theories exclude the possibility of such an interaction between systems (Sloman, 1996; Mukherjee, 2010), so this evidence does not bear on these theories. The interactive theories discussed in Chapter 1 do not include inhibition as a central role for system 2, so the findings in Chapter 2 provide no information about these theories. Chapter 3, on the other hand, provided clear evidence against MDSM, as stated by Mukherjee (2010). The fact that the model fit adequately when the function of the affective system was adjusted, though, showed that parallel-competitive models of decision making (e.g., Epstein, 1994; Sloman, 1996) can
be used to account for the effects of working-memory load on risky choice. These
findings did not test and, thus, did not rule out the possibility that default-interventionist
or interactive models can also account for the results of the experiment.

On the one hand, these findings show that dual-system approaches have at times
incorrectly characterized decision-making processes, for example through the use of
correlations with the CRT as evidence for causal claims, or by hypothesizing specific, but
incorrect mechanisms for system 1’s behavior. On the other hand, they show that some of
the blanket criticisms of these theories (see Table 2) are inaccurate, even with respect to
some verbal default-interventionist theories.

The first three criticisms (that two systems are too many, too few, or that dual-system
models are inconsistent) are addressed in the Chapter 1 and are not the subject of
empirical study in Chapters 2 and 3. Both Chapters 2 and 3, however, do address the
fourth criticism, that dual-system models do not make empirical predictions. The studies
in Chapters 2 and 3 falsified empirical predictions based on both verbal and mathematical
dual-system models, demonstrating that they can make falsifiable and precise (in the case
of mathematical models) predictions and therefore, that progress in developing better and
more accurate dual-system models is possible. Chapter 3 addressed the last criticism, that
dual-system models are not clear or specific enough. Results of that chapter demonstrated
that, even though one formulation of a parallel-competitive dual-system model was
incorrect in a very clear and specific way, another formulation could account for this
error. Although Chapter 3 addresses this criticism for MDSM, it does little to address this
criticism with respect to verbal dual-system models, which may still be considered vague about their structures by the authors that levied those criticisms. For example, Kahneman and Frederick’s (2002) model might be considered vague because it suggests that deliberative processing follows intuitive processing, but also that the two ultimately compete for the response. It is therefore vague about whether or not system 1 processing concludes by the time system 2 is active. Nonetheless, it shows that blanket criticisms of all dual-system theories as not clear or specific enough are inaccurate.

**Generalizations to Other Decision Tasks**

The analytic tools described in Chapters 2 and 3 (IRTrees and MDSM) may be able to be generalized to other decisions to test dual-system predictions. IRTrees, for example, can be applied to modeling decision problems directly, rather than to modeling CRT responses, as long as the responses can be classified into several categories that arise from intuition, inhibition of intuition, and further deliberation.

For example, the Decision Making Competence scale (Bruine de Bruin, Parker & Fischhoff, 2007) includes decision problems like the traditional Asian Disease problem which may be modeled using IRTrees. Imagine such a problem with two options, one of them may be attractive to the affective system: save 150 people for certain, and the other to the deliberative system: take a 1/3 chance to save 700 people (expected value = 233 people). The addition of a third option, which is not intuitive and is worse than the deliberative option in terms of expected value (e.g., 1/4 chance to save 800 people; expected value = 200 people) may clarify processes. Since acting consistently with
expected value requires more than foregoing unfavorable certain outcomes, at least some people who would have otherwise chosen the highest expected value option (1/3 chance to save 700) might be expected to choose this third option, indicating that they are not expected value responders, and they did not fall prey to the certainty bias that attracts the affective system.

Similarly, MDSM’s approach may be generalizable to decision problems other than choices between gambles, as long as they have features that can be systematically manipulated and the affective and deliberative systems’ roles in the decision can be clearly stated. Application to the Asian Disease problem is straightforward: Numbers of people, like quantities of money in the risky choice problems, can be valued linearly by system 2 and in accordance to a power function by system 1. Probability can be weighted linearly by system 2, and perhaps completely ignored by system 1 (alternatively, the formulation of the intuitive system that ignores lives and attends to the probabilities of saving anyone at all may be more effective).

**Difficulties in Applying Computational Methods to Interactive Models**

Many dual-system models, especially the newer models, are interactive in nature. For example, some researchers argue that emotions can produce deliberative responses and deliberation can produce emotional responses (Peters, 2006; Lerner & Keltner, 2000). Although experimental, self-report, and physiological measures of emotion can all find influences of emotional and deliberative processes on each other, it is more difficult to produce and find evidence for a precise mathematical model of the sequence of these
influences. For example, Loewenstein et al. (2001) hypothesized that anticipated outcomes and subjective probabilities lead to a cognitive evaluation, which produces both the decision and other feelings. Other researchers have emphasized the primacy of affect in the evaluative process, with a combined affective evaluation guiding deliberation (e.g., Epstein, 1994). It may be possible to make these into computational time-series models and test which of these two models is preferable, but it is necessary to have on-line process measures of both affect and deliberation. Without this type of measurement, the two types of models cannot be differentiated because the only difference between them is the order of events.

**Relation to Affect and Cognitive Heuristics**

In many dual-system models, affect plays a vital role in the operation of one or both systems (e.g., Epstein, 1994; Slovic et al., 2002; Damasio, 1994). There appears to be strong evidence, both behavioral and neuroscientific, that affect does play a role in judgments and decisions (De Martino et al., 2006; Bechara, Damasio, & Damasio, 2000). These revelations of the importance of affect require the revision of some verbal dual-system models (e.g., Kahneman, 2003; Sloman, 1996) because these models describe the cognitive processes involved in decision making in detail and largely omit the role of affect. Mathematical models, on the other hand, may or may not need revision, depending on whether their components correspond to cognitive heuristics and/or affective processes, or are neutral in terms of content. In general, IRTrees necessitate a step-wise decision process and are appropriate for modeling a default-interventionist process, but
are mute on whether it is affect or a cognitive heuristic that is being inhibited, so they need no revision. However, the application of IRTrees to the CRT in Chapter 2 did assume that it was an intuitive, not affective, process being inhibited, so if future research finds an emotional component in the CRT, the verbal description of what the IRTree represents may need revision. MDSM uses affect as the primary motivation behind system 1, so it is compatible with a vital role of affect. However, given a different description, its equations can just as easily represent cognitive heuristics because only one parameter is responsible for the degree of deliberativeness.

**Limitations of Repeated Measures Methodologies**

Both Chapters 2 and 3 used repeated measures to measure psychological parameters at the level of the individual. Psychologists are usually interested in making theories about the individual. Although many psychological theories are constructed on the basis of between-subject data, repeated measurements on the level of the individual are helpful to inform some of these types of theories because not having individual-level measurement can result in errors of aggregation discussed below.

Repeated measures are also helpful because they might approximate some real choice environments. Indeed, many choices are made multiple times in the real world (e.g., depositing money to a savings account, choosing apples at a supermarket, detecting spam emails). Inferences from experiments involving repeated measures like those conducted in Chapters 2 and 3 are more likely to be valid for these types of decisions.
However, many important decisions are made once or only a handful of times (choice of job, choice of spouse, choice of retirement option). Making some type of decision many times can influence the underlying decision processes and change decision outcomes when compared to these one-shot choices. Such changes can happen because participants become bored or inattentive after many similar choices, because they make new comparisons after being asked similar questions (Ariely, Loewenstein, & Prelec, 2006). Therefore, inferences from paradigms, such as those employed in Chapters 2 and 3, may be less valid for choices that are made only once in the real world.

One-shot decisions can also be studied using the types of models discussed. To study these types of decisions, manipulations that were made within subjects (e.g., amount of money and probability of winning in Chapter 3) can be made between subjects. But these between-subject designs come with their own difficulties. First, many subjects are needed to achieve the same degree of uncertainty in parameter estimation as was possible with only a few subjects employing repeated. Second, measurement at the level of the individual has to be sacrificed because noise at the trial level cannot be distinguished from individual differences. For example, in the case of Chapter 3, parameters $\alpha$, $\delta$, and $\gamma$ would only be measured separately for those in the working-memory load and control conditions, but not for each individual if a between-subject design was used.
Aggregation Errors

A fully between-subject design which forbids measurement of parameters on the level of the individual can be problematic because sacrificing individual-level measurement can result in a range of aggregation errors, such that inferences on the group level will not reflect an accurate inference about any individual (James, 1982). One famous example of incorrect inferences being drawn if within-subject measurements are unavailable is Simpson’s Paradox (Blyth, 1972). A famous dataset showed that women were being accepted at lower rates than men to a competitive university, apparently indicating discrimination (Bickel, Hammel, & O’Connell, 1975). However, upon closer inspection, the data indicated that the lower acceptance rates were due to women applying to more competitive departments. It was possible to ascertain this only because data were available for multiple men and women per department. Had there been no repeated measures on the level of the department in this study, it would only be possible to come away with the wrong conclusion.

Experiments with subjective evaluations as dependent variables are even more vulnerable to these types of inaccurate inferences. For example, Birnbaum (1999) showed that a group of participants asked how big the number 9 is, gave larger assessments than others who were asked how big the number 221 is, inviting the conclusion that they believe $9 > 221$, a conclusion that is obviously invalid for any one individual that might be asked that question. Of course, this study itself is a between-subject study, and provided
valuable information about the way people make subjective assessments of the sizes of numbers, so this approach is not without merit.

**Additional Limitations**

In addition to the limitations of the IRTrees approach discussed in Chapter 2, applications of IRTrees require one to assume a cognitive process that generates the observed responses. However, this assumption need not go untested. In the case of the CRT, for example, it is possible that there is only one skill responsible for performance and all types of responses (correct, intuitive incorrect, and non-intuitive incorrect) are generated by different levels of that one skill. In other words, a single factor could exist that combines Calculation and Cognitive Reflection such that a low score for an individual would indicate that most responses were non-intuitive incorrect, a medium score would indicate that most responses were intuitive incorrect, and a high score would indicate mostly correct responses. Another possibility is that responses might be generated in two steps in a different order than what was assumed in the present paper. For example, perhaps one factor decides if the response is correct, then if incorrect, a second factor decides if the error is intuitive or not. With the CRT, the two-step structure described in Chapter 2 appears most intuitive, and this structure has also empirically been tested against the alternatives described above (the graded response model and the two steps in the opposite order) by Böckenholt (2012b). In the case of other decision problems and biases, more work needs to be done to determine if it is valid to assume intuition inhibition is the data generating process.
MDSM also has several limitations not discussed thus far. Similar to IRTrees, MDSM assumes a structure in which the two systems do not interact. Like IRTrees, it can easily be tested against alternative structures, as long as they can also be expressed mathematically and fit to data. A more serious flaw is that the model does not specify exactly what circumstances change the key $\gamma$ parameter (the weight of the affective value of the gamble), except for two cases: 1) with ambiguous options, $\gamma$ is expected to be higher and 2) as in the present Chapter 3, $\gamma$ increases with greater affect and less deliberation. The experimental manipulations that may result in greater affect or less deliberation are left to be determined by other theories. On the one hand, this vagueness invites criticisms of vagueness discussed with respect to other dual-system theories. Specific manipulations used in previous literature (e.g., incentivized outcomes or greater amounts to win/lose) might result in greater affect or greater deliberation, predicting a change in $\gamma$ in either direction or not affecting it at all. On the other hand, if the theory is correct, it reduces this vagueness to the value of a parameter and it allows affectiveness to be measured in the choice context using that parameter.
REFERENCES


Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive


Behavioral Decision Making, 27(2), 95-108.


McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from
the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419-457.


APPENDIX A: ROBUST REGRESSIONS

Robust regressions were conducted in order to ensure that the way we chose inattentive or outlier participants did not influence results. The results of these regressions are comparable to those of regular regression, except they are much less susceptible to be influenced by a minority of observations that differ from the general trend. These models are identical to regression models at the first iteration. Observations are iteratively reweighted according to their distance from the estimated regression line, with further observations being underweighted. The process is repeated until convergence (Holland & Welsch, 1977). Robust regressions were implemented through the MASS package in R (Venables & Ripley, 2002). Below are results reported in Tables 7 and 9. Note that demographic variables were also controlled for in this version of Table 7, but are not displayed for simplicity.
Table 11. Robust regression results for Chapter 1 Study 1

<table>
<thead>
<tr>
<th></th>
<th>Frame Inconsistency</th>
<th>Conjunction (subset vs. superset)</th>
<th>Conjunction (time)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without Numeracy</td>
<td>With Numeracy</td>
<td>Without Numeracy</td>
</tr>
<tr>
<td>Intercept</td>
<td>2.11</td>
<td>2.01</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>(0.20)</td>
<td>(0.20)</td>
<td>(0.10)</td>
</tr>
<tr>
<td>Cognitive Reflection</td>
<td>-0.06</td>
<td>-0.05</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>(0.05)</td>
<td>(0.05)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Calculation</td>
<td>-0.13</td>
<td>-0.10</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.04)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>Numeracy</td>
<td>--</td>
<td>-0.11</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>(0.04)</td>
<td>(0.02)</td>
<td></td>
</tr>
</tbody>
</table>
Table 12. Robust regression results for Chapter 1 Study 2

<table>
<thead>
<tr>
<th>DV</th>
<th>Intercept</th>
<th>Education</th>
<th>Income</th>
<th>Age*</th>
<th>Gender</th>
<th>Intel</th>
<th>Cog Refl</th>
<th>Calculation</th>
<th>Numeracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Biases</td>
<td>1 0.06 (-0.14)</td>
<td>-0.01 (0.01)</td>
<td>-0.01 (0.01)</td>
<td>0.01 (0.01)</td>
<td>0.05 (0.03)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 0.00 (0.14)</td>
<td>-0.00 (0.01)</td>
<td>-0.01 (0.01)</td>
<td>0.01 (0.01)</td>
<td>0.05 (0.03)</td>
<td>-0.20 (0.03)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 0.01 (0.14)</td>
<td>-0.00 (0.01)</td>
<td>-0.00 (0.01)</td>
<td>0.01 (0.01)</td>
<td>0.03 (0.03)</td>
<td>-0.20 (0.04)</td>
<td>-0.10 (0.05)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 0.03 (0.14)</td>
<td>0.00 (0.01)</td>
<td>-0.00 (0.01)</td>
<td>0.01 (0.01)</td>
<td>0.02 (0.03)</td>
<td>-0.17 (0.03)</td>
<td>0.02 (0.06)</td>
<td>-0.16 (0.05)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 0.20 (0.14)</td>
<td>0.01 (0.01)</td>
<td>-0.00 (0.01)</td>
<td>0.00 (0.01)</td>
<td>0.01 (0.03)</td>
<td>-0.13 (0.04)</td>
<td>0.06 (0.06)</td>
<td>-0.10 (0.05)</td>
<td>-0.45 (0.10)</td>
</tr>
<tr>
<td>Financial Outcomes</td>
<td>1 0.62 (0.04)</td>
<td>0.007 (0.003)</td>
<td>0.004 (0.002)</td>
<td>0.024 (0.004)</td>
<td>-0.02 (0.01)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 0.64 (0.04)</td>
<td>0.005 (0.003)</td>
<td>0.003 (0.002)</td>
<td>0.025 (0.004)</td>
<td>-0.03 (0.01)</td>
<td>0.04 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 0.64 (0.04)</td>
<td>0.004 (0.003)</td>
<td>0.003 (0.002)</td>
<td>0.024 (0.004)</td>
<td>-0.02 (0.01)</td>
<td>0.04 (0.01)</td>
<td>0.02 (0.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 0.64 (0.04)</td>
<td>0.004 (0.003)</td>
<td>0.003 (0.002)</td>
<td>0.024 (0.004)</td>
<td>-0.02 (0.01)</td>
<td>0.04 (0.01)</td>
<td>0.02 (0.02)</td>
<td>0.01 (0.01)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5 0.60 (0.04)</td>
<td>0.003 (0.002)</td>
<td>0.003 (0.002)</td>
<td>0.026 (0.004)</td>
<td>-0.02 (0.01)</td>
<td>0.03 (0.01)</td>
<td>0.01 (0.02)</td>
<td>-0.00 (0.01)</td>
<td>0.09 (0.03)</td>
</tr>
</tbody>
</table>
APPENDIX B: MDSM SIMULATIONS

First, \( \gamma, \alpha, k \) were sampled using a random uniform distribution. Parameters \( \gamma \) and \( k \) were sampled 100 times on the range between 0 and 1 and \( \alpha \) was sampled between 0 and 1.1. Simulations of 100 experiments similar to that described in Chapter 3 were conducted. Probability of choice was assumed to be a logistic function of the difference between values calculated through MDSM. The model was fit to each of the 100 simulated experiments to see if it could recover the parameters that were used to generate the choices in each case. Because the logit function resulted in some noise added to the responses, parameters could not be expected to be recovered perfectly. However, since the sample size was sufficient, estimates should approach simulated values.

When \( k \) was left as a free parameter, \( \alpha \) was difficult to estimate with RMSEA of .17. This was especially true for low values of \( \gamma \). However, when \( k \) was fixed to \( k = 3^{\alpha-1} \), like in Chapter 3, \( \alpha \) and \( \gamma \) were estimated successfully with RMSEAs of .05 and .03 respectively. This was not much different from Prospect Theory parameters estimated on the same dataset, which had RMSEAs of .004 and .03.

However, in many model fits reported in Chapter 3, there was an additional parameter: \( \delta \). Simulations showed that in this case, again if \( \gamma \) was low (<.2), the model had trouble recovering parameters, but when it was high, the model did fine. Overall, the model
parameters $\gamma, \alpha, \delta$ were recovered with moderate success with RMSEAs of .11, .04, and .15 respectively. These were not much different from when prospect theory was fit with a logit parameter (a common practice). Even the simplest possible 2-parameter Prospect Theory had RMSEAs of .09 and .07 for its parameters.

If all four quantities, $\gamma, \alpha, \delta, k$ would be left to be estimated, however, the model would not be identified. The model is not identifiable with $k$ left to be estimated for the following reason.

Consider the scaled difference in value in the model (the input to the logit function):

$$\delta(V(G_2) - V(G_1)) = \delta(\gamma V_A(G_2) + (1 - \gamma)V_D(G_2) - \gamma V_A(G_1) + (1 - \gamma)V_D(G_1))$$

It can be re-written as follows:

$$\delta(V(G_2) - V(G_1)) = \delta(\gamma V_A(G_2) - V_A(G_1)) + \delta(1 - \gamma)(V_D(G_2) - V_D(G_1))$$

$$\delta(V(G_2) - V(G_1)) = \delta\gamma (V_A(G_2) - V_A(G_1)) + \delta(1 - \gamma)k\left(\sum_{i=1}^{n} p_{2,i}x_{2,i} - \sum_{i=1}^{n} p_{1,i}x_{1,i}\right)$$

However, there are only two quantities ($f_1$ and $f_2$ below) that vary as a function of the data and three parameters to be estimated:

$$\delta(V(G_2) - V(G_1)) = \delta f_1(D) + \delta(1 - \gamma)k f_2(D)$$

Therefore, the system is undertermined.
APPENDIX C: MDSM FITS WITHOUT DELTA

A simple version of MDSM without the $\delta$ parameter was fit to data. The model fits to data are shown below. This model was able to capture some of the differences between conditions in its parameters, as you can see by the diverging black and gray lines. However, the model provided missfits of behavior for the $1$ for sure vs. gamble choices in addition to the misfits of the regular MDSM in the gamble vs. gamble choices.
Figure 12. MDSM fits without delta to proportions choosing certain option
Figure 13. MDSM fits without delta to proportions choosing higher likelihood gamble
APPENDIX D: STAN CODE FOR MDSM

data {
  int<lower=1> T;
  int<lower=1> N;
  int<lower=1> subj[T];
  int<lower=0,upper=1> wmcond[N];
  real<lower=0,upper=1> leftp[T];
  int<lower=0,upper=10> leftd[T];
  real<lower=0,upper=1> rightp[T];
  int<lower=0,upper=10> rightd[T];
  int <lower=0,upper=1> response[T];
}

parameters {
  real<lower=0, upper=1000> hyper_mu_d;
  real<lower=0, upper=1000> hyper_mu_g;
  real<lower=0, upper=1000> hyper_sigma_d;
  real<lower=0, upper=1000> hyper_sigma_g;

  vector<lower=0, upper=1000>[3] mu;
  vector<lower=0, upper=1000>[3] sigma;

  vector<lower=0, upper=1000>[2] mu_wm;
  vector<lower=0, upper=1000>[2] sigma_wm;

  vector<lower=0, upper=1000>[N] alpha;
  vector<lower=0,upper=1>[N] gamm;
  vector<lower=0, upper=1000>[N] error;
  #vector<lower=0, upper=1000>[N] k;
}

model {

}
vector[T] lvalue;
vector[T] rvalue;
vector[T] choicep;
vector[N] k;

for (i in 1:N) {
    k[i] <- 3^(alpha[i]-1);
}

for (t in 1:T) {
    if(leftp[t] > .99) {
        lvalue[t] <- gamm[subj[t]]*leftd[t]^alpha[subj[t]] + 
                    (1-gamm[subj[t]])*k[subj[t]]*leftd[t];
    } else {
        lvalue[t] <- gamm[subj[t]]*0.5*leftd[t]^alpha[subj[t]] + 
                    (1-gamm[subj[t]])*k[subj[t]]*leftd[t]*leftp[t];
    }

    if(rightp[t] > .99) {
        rvalue[t] <- gamm[subj[t]]*rightd[t]^alpha[subj[t]] + 
                    (1-gamm[subj[t]])*k[subj[t]]*rightd[t];
    } else {
        rvalue[t] <- gamm[subj[t]]*0.5*rightd[t]^alpha[subj[t]] + 
                    (1-gamm[subj[t]])*k[subj[t]]*rightd[t]*rightp[t];
    }

    choicep[t] <- 1/ (1.000001 + exp(-error[subj[t]]* 
                        (lvalue[t]-rvalue[t])));
}

for (n in 1:N) {
    gamm[n] ~ beta(mu[2]*sigma[2]*(1-wmcond[n]) + 
                    mu_wm[1]*sigma_wm[1]*wmcond[n], 
                    178}
(1-mu[2])*sigma[2]*(1-wmcond[n]) + (1-mu_wm[1])*sigma_wm[1]*wmcond[n]);
          mu[3]/sigma[3]^2*(1-wmcond[n]) +
          mu_wm[2]/sigma_wm[2]^2*wmcond[n]);
}

mu[1] ~ cauchy(0.5,1);
mu[2] ~ beta(hyper_mu_g*hyper_sigma_g, (1-hyper_mu_g)*hyper_sigma_g);
mu[3] ~ cauchy(hyper_mu_d, hyper_sigma_d);
#mu[4] ~ cauchy(1,10);
sigma[1] ~ cauchy(1,2);
sigma[2] ~ cauchy(10,20);
sigma[3] ~ cauchy(50, 150);
#sigma[4] ~ cauchy(20,75);

mu_wm[1] ~ beta(hyper_mu_g*hyper_sigma_g, (1-
          hyper_mu_g)*hyper_sigma_g);
mu_wm[2] ~ cauchy(hyper_mu_d, hyper_sigma_d);
sigma_wm[1] ~ cauchy(10,20);
sigma_wm[2] ~ cauchy(50, 150);

hyper_mu_g ~ uniform(0,1);
hyper_mu_d ~ cauchy(4,20);
hyper_sigma_g ~ cauchy(3, 20);
hyper_sigma_d ~ cauchy(10, 50);

response ~ bernoulli(choicep);
APPENDIX E: EXPECTED VALUE MODEL FITS

An expected value model with only one parameter to estimate ($\delta$) was fit to data. The model fits to data are shown below. The fits were very similar to those of MDSM without $\delta$. This model was able to capture some of the differences between conditions in its parameters, as you can see by the diverging black and gray lines. But like the simple MDSM, the model provided gross missfits of behavior both in the $1 for sure vs. gamble and gamble vs. gamble choices as shown below.
Figure 14. Expected value model fits to proportion choosing certain option
Figure 15. Expected value model fits to proportions choosing higher likelihood gamble