A DIFFUSION MODEL ANALYSIS OF TRANSITIVITY AND LEXICOGRAPHIC SEMIORDER

A Dissertation presented to
the Faculty of the Graduate School
at the University of Missouri-Columbia

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by

SANGHYUK PARK

Dr. Clintin Davis-Stober, Dissertation Supervisor

MAY 2023

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled:

A DIFFUSION MODEL ANALYSIS OF
TRANSITIVITY AND LEXICOGRAPHIC SEMIORDER

presented by Sanghyuk Park,

a candidate for the degree of Doctor of Philosophy in Psychology,

and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Clintin Davis-Stober

Professor Erin Schliep

Professor Ed Merkle

Professor Nelson Cowan

ACKNOWLEDGMENTS

I thank my family, my wife Autumn, my daughters Olivia and Sophia, and my son Ethan, but especially Autumn. This dissertation couldn't have been finished if I had to work alone without her support. There had been ups and downs during the process, but she had always been there for me, encouraging me to take courage when the task felt too daunting. She is a strong person, and I appreciate every moment I spent with her.

I thank Clintin Davis-Stober for being my dissertation advisor and being my mentor for the last 10 years. Clint has been a great advisor I could ever ask for. Every time I had a meeting with him, he provided me with great insights and helped me see the task from a completely different perspective. Specifically, he helped me look at the big picture of the current dissertation, and it helped immensely in deciding every detail of the dissertation. I will never forget the last 10 years, especially the first 6 years, in which I was able to work with him. It was truly an amazing opportunity.

I thank Dr. Merkle, Dr. Cowan, and Dr. Schliep for being in my committee members and for their great insights into my research. Their expertise in their respective fields have helped me enlarge my understanding about my work. I am very grateful for all the comments and valuable suggestions they have made for me.

Lastly, but not the least, I thank my family in South Korea and my family-in-law for their love, patience, support, and encouragements. Because of them, I was able to pursue the dissertation to the end even when I felt like giving up. Thank you.

TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . . **ii**

LIST OF TABLES . . . . **vi**

LIST OF FIGURES . . . . **vii**

LIST OF ABBREVIATIONS . . . . **xii**

ABSTRACT . . . . **xiii**

Chapter . . . . **xiii**

1 Introduction **1**

1.1 Research questions 4

1.2 Roadmap of dissertation 7

2 Transitivity and lexicographic semiorder **9**

2.1 Transitivity and lexicographic semiorders 13

2.1.1 Transitivity and rationality 13

2.1.2 Lexicographic Semiorders and Heuristics 18

2.2 Statistical models for decision theories 25

2.2.1 Variability as error 26

2.2.2 Random preference 33

2.2.3 Statistical model of axiomatic representational theories 35

2.3 Statistical inferences 49

2.3.1 Frequentist statistical inference 50

2.3.2 Bayesian inferences 54

3 Diffusion model **69**

3.1 Diffusion model 72

3.1.1 Sequential sampling framework 72

3.1.2 Parameters of the diffusion model 78

3.1.3 Mathematical formulation of the diffusion model 82

3.2 Applying the diffusion model to preferential choice tasks 85

3.2.1 Background 85

3.2.2 The data to be examined 87

3.3 Model specification 88

3.3.1 Parameter constraints 88

3.3.2 Priors for parameters 92

4 Simulation study: Parameter recovery simulations 102

4.1 Background 103

4.1.1 Diffusion model 103

4.1.2 Preferential choices 104

4.2 Simulating data 106

4.2.1 Simulating data from the full diffusion model 107

4.2.2 Simulating data from the constrained diffusion model 111

4.3 Model specification 112

4.4 Results 119

4.4.1 Result: The constrained data 120

4.4.2 Result: The full data 129

4.5 Discussion 133

5 Empirical analysis: Application to real data 137

5.1 Cavagnaro and Davis-Stober's (2014) gamble experiment 139

5.1.1 Data set 139

5.1.2 Results 141

5.2 In-lab experiment 159

5.2.1 Data set 159

5.2.2 Results 162

5.3 Online experiment 168

5.3.1 Data set 168

5.3.2 Results 171

5.4 Discussion 176

6 General discussion 178

6.1 Back to research questions 181

6.1.1 How do choices differ? 181

6.1.2 Why do choices differ? 183

6.2 Limitations and future research 184

APPENDIX 185

A Stan code for the diffusion model analysis 185

References 189

VITA 209

LIST OF TABLES

Table Page

2.1 Full list of lexicographic semiorders over $\{A,\ B,\ C\}$ 41

4.1 Group-level parameters 109

4.2 Individual-level parameters 110

4.3 Bayesian p-values for simulation study 124

4.4 Observed and predicted choice proportions for Participant 14 125

4.5 Estimates for group-level parameters of the constrained diffusion model 128

4.6 Estimates for group-level parameters of the full diffusion model 133

5.1 Gamble sets 140

5.2 Decision criteria for Bayes factor 143

5.3 Bayes factors of Cavagnaro & Davis-Stober (2014) 145

5.4 Summary of classifications of Cavagnaro & Davis-Stober (2014) . . . 145

5.5 Mean response time for gamble pairs between WST and LSEM 148

5.6 MCMC sampling information 151

5.7 Bayesian p-values of Cavagnaro & Davis-Stober (2014) 152

5.8 Posterior estimates of μ_a^{WST} and μ_a^{LSEM} 157

5.9 Bayes factors of in-lab data 163

5.10 Summary of classifications of in-lab data 163

5.11 Bayesian p-values of in-lab data 165

5.12 Summary of classifications for online data 172

LIST OF FIGURES

Figure Page

2.1 Gamble set employed in Tversky's (1969) experiment. Gambles are in a pie format, where the blue-colored area represents the probability of winning the specified payoff. The expected values of the gambles increase with the probability of winning. 22

3.1 A basic framework of the diffusion model. a represents the decision criterion, z represents the starting point of the accumulation process, v represents the drift rate, and t_{er} represents the non-decision time. When the accumulator reaches one of decision criteria (either 0 or a), the decision is made. 79

4.1 Gamble set used to simulate the data. Gambles are in the pie format, where the blue-colored area represents the probability of winning the specified payoff. Different gambles have different values of payoff and probability of winning, in a way that the two attributes tradeoff with each other. 108

4.2 Prior predictive distributions of a for conservative and less-conservative groups. Majority of probability densities of both distributions are centered around 1. 115

4.3 Cumulative density function for the diffusion model. To generate this CDF, the main parameters of the diffusion model are set as follows: $a = 2$, $t_{er} = 0.3$, $z = 0.5$, and $v = 1$. The CDF approaches toward its limiting value of 0.88 as response time increases. 122

4.4 Density plots of response time for the full and constrained data. The RT data are collapsed across participants. The density of the observed RT is displayed via a histogram, and the density of the predicted RT is displayed via a density curve. 126

4.5 Density plots of response time of each participant for the constrained data. The density of the observed RT is displayed via a histogram, and the density of the predicted RT is displayed via a density curve. . 127

4.6 Posterior estimates of decision criterion parameter a for the constrained data. Participants are arranged in an ascending order, based on the estimates of a. Empty circled points represent posterior mean, line segments represent the 95% credible intervals, and filled circled points represent true values. Different colors are used to indicate different classifications of decision-making styles, with vertical lines representing the estimates of the group-level decision criterion μ_a for each classification. 129

4.7 Posterior estimates of subjective values $\boldsymbol{u}$ for the constrained data. Empty circled points represent posterior mean, line segments represent the 95% credible intervals, and filled circled points represent true values. Since the subjective values are centered with respect to Gamble C, all subjective values are estimated relative to Gamble C, with Gamble C's subjective value fixed at 0. 130

4.8 Density plots of response time of each participant for the full data. The density of the observed RT is displayed via a histogram, and the density of the predicted RT is displayed via a density curve. 132

4.9 Posterior estimates of decision criterion parameter a for the full data. Participants are arranged in an ascending order, based on the estimates of a. Empty circled points represent posterior mean, line segments represent the 95% credible intervals, and filled circled points represent true values. Different colors are used to indicate different classifications of decision-making styles, with vertical lines representing the estimates of the group-level decision criterion μ_a for each classification. 134

4.10 Posterior estimates of subjective values $\boldsymbol{u}$ for the full data. Empty circled points represent posterior mean, line segments represent the 95% credible intervals, and filled circled points represent true values. Since the subjective values are centered with respect to Gamble C, all subjective values are estimated relative to Gamble C, with Gamble C's subjective value fixed at 0. 135

5.1 Gamble sets 1 and 2 from Cavagnaro and Davis-Stober (2014). Set 1 consists of gambles that have expected value increasing with probability winning. Set 2 consists of gambles that have expected value increasing with payoff. 139

5.2 Density plot of response time for Gamble Set 1 from Cavagnaro & Davis-Stober (2014). The observed response times are plotted as a histogram, and the predicted response times are plotted as a density curve superimposed on the histogram. Response times are plotted separately for different classifications. 153

5.3 Density plot of response time for Gamble Set 2 from Cavagnaro & Davis-Stober (2014). The observed response times are plotted as a histogram, and the predicted response times are plotted as a density curve superimposed on the histogram. Response times are plotted separately for different classifications. 154

5.4 Posterior estimates of subjective values for Gamble Set 1 from Cavagnaro & Davis-Stober (2014). Filled dots indicate posterior mean of subjective values of each gamble, and vertical lines overlaid on the filled dots indicate the 95% credible intervals. 155

5.5 Posterior estimates of subjective values for Gamble Set 2 from Cavagnaro & Davis-Stober (2014). Filled dots indicate posterior mean of subjective values of each gamble, and vertical lines overlaid on the filled dots indicate the 95% credible intervals. 156

5.6 An example trial of the new experiment. Now participants get to play the chosen gamble every trial, and the money they win is accumulated in their virtual account. How much money they have earned is shown in the middle on screen. 160

5.7 An example of the chosen gamble being played. If the arrow points to the blue-colored area, participants win the specified amount of money. Otherwise, they win nothing. 160

5.8 Gamble Set 3. All the gambles in Set 3 have the same expected monetary value. . 162

5.9 Density plots of RT for Gamble Set 1 in in-lab data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve. 166

5.10 Density plots of RT for Gamble Set 2 in in-lab data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve. 166

5.11 Density plots of RT for Gamble Set 3 in in-lab data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve. 167

5.12 Subjective values of Gambles in Set 1 in in-lab data. Filled dots indicate posterior mean, and vertical lines superimposed on the dots indicate the 95% credible intervals. 168

5.13 Subjective values of Gambles in Set 2 in in-lab data. Filled dots indicate posterior mean, and vertical lines superimposed on the dots indicate the 95% credible intervals. 169

5.14 Subjective values of Gambles in Set 3 in in-lab data. Filled dots indicate posterior mean, and vertical lines superimposed on the dots indicate the 95% credible intervals. 170

5.15 Histogram of observed RT from the online experiment. The vertical line indicates a cut-off value of 0.3 we adapted for other data sets. 171

5.16 Density plots of RT for Gamble Set 1 in online data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve. 174

5.17 Density plots of RT for Gamble Set 2 in online data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve. 175

5.18 Density plots of RT for Gamble Set 3 in online data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve. 176

LIST OF ABBREVIATIONS

2AFC	2-Alternative Forced Choice
BF	Bayes Factor
CDF	Cumulative Density Function
LSEM	Lexicographic Semiorder Error Model
MCMC	Markov Chain Monte Carlo
MVN	Multi-Variate Normal distribution
PDF	Probability Density Function
RT	Response Time
WST	Weak Stochastic Transitivity

ABSTRACT

The primary aim of the present dissertation is to examine the underlying cognitive processes of transitivity and lexicographic semiorders. To this end, I apply the diffusion model to preferential choice data, where transitivity or lexicographic semiorders are typically considered to model the choice data. In literature, transitivity is often associated with rationality, whereas lexicographic semiorders are usually considered an alternative way to make decisions specifically when the given task seems daunting. Despite their clearly different decision-making processes, little empirical evidence of such different cognitive processes has been reported, so I decide to run the diffusion model analysis to provide empirical evidence of the underlying cognitive processes behind these two models. To do that, I reparameterize drift rate of the diffusion model in terms of subjective values (or utility) of the alternatives. And I conduct a simulation study to test the new diffusion model's ability to recover the data-generating parameter values. Then, I apply the diffusion model to three sets of real data, one from Cavagnaro and Davis-Stober's (2014) experiment, and two from my own experiment. The results imply that people classified to transitivity tend to integrate more information than those classified to lexicographic semiorders to make a decision. More details about the results and implications are discussed in Chapters 4 and 5.

Chapter 1

Introduction

Suppose that you are an undergraduate student in psychology, and to fulfill one of the requirements of psychology courses, you participate in a decision-making experiment. You are now in one of the rooms where the experiment is administered and asked to answer every decision task that will appear on screen. The task is to choose between two monetary gambles (or lotteries). Each gamble is composed of two outcomes: winning some amount of money and winning nothing. And each outcome is associated with its own probability. Different gambles are defined with different amounts of payoff and different probabilities of winning. Undoubtedly, you find that high payoff is coupled with low probability of winning and low payoff is coupled with high probability of winning. That is, you cannot pursue both attributes, i.e., payoff and probability of winning, at the same time; you must make a trade-off. In this case, what would you choose?

The experimental paradigm illustrated above describes how a typical decision making experiment is administered in psychology and economics. Every participant in such experiment is confronted with a series of binary choice tasks, and researchers are interested in learning how people answer those tasks. Interestingly, even though the task is same for every participant in the experiment, the responses vary from individual to individual. Such individual difference may seem natural, because everyone thinks differently. But, if we think over the different responses, it raises some important questions. Do the responses differ in a haphazard manner, or do they differ systematically? What makes them respond differently, or why do they differ?

The current study is an endeavor to tackle the questions above. The questions comprise two main parts: how the responses differ, and why the responses differ. The first part of the question addresses the "how" aspect of the individual differences, concerned with different patterns of choices. In other words, this is the problem of modeling different patterns of preferences that give rise to the choices we observe. At the level of decision task, it may seem like one makes whichever choices he or she likes to make for the given task; but once we take a step back and look at the data at the level of the whole experiment, a choice for each task begins to make a sequence of choices, where clearly different choice patterns emerge across individuals. Those different patterns of choices are what psychologists and economists have long endeavored to model, and the particular model I choose is the representational theories of preferences. Specifically, I consider two of the most renowned representational theories: transitivity and lexicographic semiorders.

The second part of the question addresses the "why" aspect of the individual differences, concerned with the cognitive mechanisms that underlie the different observed responses. In order to examine the underlying cognitive processes, one should consider incorporating response time (RT) into the analysis, because RT has been known as one of the measurements that reflect our inner mental structures (Luce, 1986). But, studying RT is not an easy task; RT is a skewed distribution, presenting an enough challenge to model observed RT on its own, but when we consider that RT behaves differently depending on its accompanying choices, modeling RT easily becomes a non-trivial task. Thus, a successful model should be well capable of dealing with skewed distributions, as well as able to analyze choices and RT within the same integrated framework. To this end, I choose the diffusion model (Ratcliff, 1978). The diffusion model is one of the most, if not the most, influential and widely studied cognitive models in psychology. As one of sequential sampling models, the diffusion model accounts for decision making processes via the information accumulation process, by analyzing RT and choices at the same time. The diffusion model has shown its usefulness in countless empirical studies, providing deep insights into how our cognitive system works when we make decisions. I believe the diffusion model particularly helps us answer the second part of the question.

The primary approach taken by the current study is to combine the aforementioned two great theories, i.e., the representational theory and diffusion model theory, in an endeavor to answer the questions listed above. Each theory has been successfully applied in numerous settings in its own field. Transitivity, for example, is a necessary condition of all utility-based economic theories and usually considered a quality of rational choices. Lexicographic semiorders, on the other hand, allow for violations of transitivity, inspiring most of heuristic-based decision making rules, which provide an alternative ways to traditional decision theories. Due to the distinct choice patterns each theory predicts, it is a common practice to employ these two theories to classify observed data. So, it is not new to use representational theories to analyze choice data. What makes the current study special from other decision-making studies is the application of diffusion model analysis to the representational theories, thereby being able to investigate the underlying cognitive processes of the decision theories. This is a novel approach at least in the field of decision theories because the diffusion model had been considered not suitable for economic choice tasks, like the one illustrated in the beginning of this chapter, until very recently. However, newly discovered findings in neurophysiology suggest that the decision-making mechanisms of economic choices also follow the general scheme of the diffusion model. Such findings have motivated me for the current dissertation.

Therefore, the major contribution of this study is to provide empirical evidence of two clearly different cognitive processes for transitivity and lexicographic semiorders. Thus far, examining the cognitive process that underlies each decision theory has relied on either theoretically motivated accounts (Gigerenzer & Goldstein, 1996), or custom designed experiments, where different cognitive processes presumably lead to unique outcomes (Rieskamp & Hoffrage, 2008). Or, more directly, researchers even interviewed some of participants in the post-experimental session to discover what processes they had actually used to arrive at the decisions they made during the experimental session (Tversky, 1969). Those approaches are not necessarily bad, but it's all indirect measures of cognitive processes, where the inferences are made based on the researchers' hypotheses on the observed responses, or participants' report. The current study takes neither of those approaches mentioned here; instead, it incorporates response time into the analysis and examines the cognitive processes via the lens of the diffusion model. The main strength of the diffusion model that other approaches lack is the ability to analyze both RT and choices simultaneously within the same framework. This is important because even if one makes the same choices as others, there might be significant difference in RT among them, which can be clear evidence for different cognitive processes. Thus, unlike previous approaches, the current study provides empirical measures of underlying cognitive processes by analyzing both RT and choices, which I believe provides more direct evidence of different cognitive processes between the two decision theories.

1.1 Research questions

The primary goal of the present dissertation centers on two big research questions as elaborated above. The first question is concerned with "how" the data differ, so for this question, I consider transitivity and lexicographic semiorders. More specifically,

- **Question 1.** Can transitivity and lexicographic semiorders account for different patterns of choices?
  - **Question 1a.** If so, how many participants can be accounted for by either transitivity or lexicographic semiorders?

The above questions directly deal with different patterns of choices. If the choices are made in a haphazard manner, neither theories should not be able to account for the data well. If the data are well supported by either transitivity or lexicographic semiorders, this provides evidence for the presence of systematic patterns in data. In that case, the data can be classified into either transitivity or lexicographic semiorders. Then, we are ready to tackle the second part of the questions, "why" the data differ. To answer this question, I run the diffusion model analysis on the data, with the resulting classification of the decision theories entering the model as a group variable. Then, the diffusion model analysis will tell us if there exists any difference in underlying cognitive process between the two theories, through the values of the parameters in the model.

- **Question 2.** Can the diffusion model shed light on the difference in underlying cognitive process between transitivity and lexicographic semiorders?
  - **Question 2a.** Is there difference in decision style (i.e., conservative or less-conservative) between transitivity and lexicographic semiorders, which manifests via the decision criterion parameter in the diffusion model?
  - **Question 2b.** Is there difference in the way participants evaluate the given alternatives between transitivity and lexicographic semiorders, which manifests via the drift rate parameter in the diffusion model?

As specified in the questions above, the diffusion model has parameters that account for specific cognitive processes. The model has four main parameters, of which the decision criterion and drift rate are of substantive interest in the present dissertation (the other two are non-decision time and starting point, which I discuss in detail in Chapter 3). The decision criterion, denoted by a, determines the amount of information one needs to accumulate for a decision; the drift rate, denoted by v, is the parameter that governs the rate of information uptake in the information accumulation process. The two parameters are typically considered the core parameters in the diffusion model because it's these two parameters that are most responsible for the whole decision-making process. In other words, different sets of values of these parameters will lead to completely different cognitive processes. Hence, I hypothesize that there are significant differences in those two parameters between transitivity and lexicographic semiorders, which can be expressed in the form of null hypotheses as follows:

- $H1_0: \ a_{\text{transitivity}} = a_{\text{lexicographic semiorders}},$
- $H2_0: \ v_{\text{transitivity}} = v_{\text{lexicographic semiorders}}.$

As will be clear in later chapters, the drift rates v will be reparameterized in terms of subjective values of the given alternatives. Thus, the second hypothesis will be tested via the subjective values.

Although the hypotheses are stated under the null hypothesis significance testing framework, the current analysis exclusively uses Bayesian statistics, where I quantify evidence against the null hypotheses using posterior density of each parameter. And testing the above hypotheses the Bayesian way is at the heart of the present dissertation.

1.2 Roadmap of dissertation

This dissertation is a journey that I take to explore the field of decision-making. On the way I study various theories to account for the patterns of choices that we observe and the cognitive processes that underlie those choices. The present study doesn't aim to cover numerous decision theories in the literature, instead it focuses on a few theories that have helped me answer the questions above, along the journey. Here is how the dissertation is organized. Chapter 2 is concerned with representational theories of preference, specifically, transitivity and lexicographic semiorders. I examine those two theories in the framework suggested by Luce and Narens (Luce & Narens, 1994), where they raise the problem of how to fill the gap between deterministic decision theories and inherently variable observed choices. Thus, the first chapter will be devoted mostly to reviewing statistical models of transitivity and lexicographic semiorders and methods of testing them.

Chapter 3 is concerned with the diffusion model. As discussed above, the application of the diffusion model to representational theories is the central thesis of the current dissertation. So, I elaborate the motivation for the use of diffusion models in the analysis of economic choice data, with the recent findings in neurophysiology that support such applications. Note, however, that the parameters in the diffusion model are originally designed for cognitive tasks, such as the moving dot task, or lexical decision task. Thus, for the diffusion model to work for a new experimental paradigm, it is often required to re-parameterize some of the parameters in terms of the new setting where the diffusion model is being applied. In the current analysis, I particularly re-parameterize drift rates in terms of economic binary choices, where the alternatives are monetary gambles. Thus, in Chapter 3, I introduce how I define drift rates, as well as other parametric assumptions that I make for the economic choice task paradigm.

In Chapter 4, I run a parameter-recovery simulation study to verify the behavior of the model I propose in Chapter 3. The diffusion model has frequently shown unreliable performance in parameter-recovery simulations. This erratic behavior occurs specifically when the diffusion model is set to its full complexity, with all parameters assumed to vary every trial. The particular model I use for the current analysis is not the full diffusion model, but with the new parameterization of drift rates, it is crucial to conduct such a parameter-recovery simulation to understand the behavior of the current model, specifically where it fails. In this simulation study, two data sets are generated: one from the full diffusion model, and one from the main model I plan to use for the analysis. I then fit the same model to both data sets to see if the model is even able to recover the data-generating parameters of the full model. I discuss the results of both data sets and suggest some advice on how to interpret results.

Then, in Chapter 5, I test the model against the existing data from a published study (Cavagnaro & Davis-Stober, 2014) to see if the hypotheses are supported with the real data. In doing so, however, I found some features of the previous experiment that might have biased participants' choices. So, I design a new experiment with some added features to resolve what I consider the source of biases, and I run the diffusion model analyses on the data from my own experiment as well.

Finally, I conclude the dissertation with the discussion chapter.

Chapter 2

Transitivity and lexicographic semiorder

We make choices every day and every moment. Those choices include trivial choices, such as what to eat for lunch, but also include non-trivial choices, such as which careers to pursue. Everyone may have a different principle to follow when in need of making choices, but most economists would agree to accept the concept of utility, with little controversy, to guide how we ought to make such decisions. Theoretically, utility is considered a measurement unit of self-interest (Mcfadden, 2001). Rational economic models are built around this theoretical concept of utility, postulating that a rational agent of economics would undoubtedly make choices in the direction of maximizing his or her self-interest, or utility (Simon, 1978). As Taussig (1912) pointed out, whenever a person makes a choice for one alternative over others, there must be a certain wish that can be gratified by that particular choice, which couldn't be satisfied by other choices. Even though such feeling of gratification may last only momentarily, one's self-interest must have been maximized at that moment by that choice. In other words, if one's choices are to be rational, one should follow the principle of maximizing utility in making choices; or the agent would risk settling for mediocre choices, with other better options out there that would provide the agent with more utility.

In economics, the principle of maximizing utility is championed by expected utility theory (von Neumann & Morgenstern, 1947). Expected utility theory is concerned with decisions under risk and guides us on how to maximize utility when the outcomes of each choice are not certain. Importantly, the expected utility theory relies upon four axioms, and these axioms are often considered qualities of a rational decision maker. The four axioms are *completeness*, *transitivity*, *independence of irrelevant alternatives*, and *continuity*. Among them, the axiom of transitivity has been particularly associated with rationality (Bar-Hillel & Margalit, 1988; Davis-Stober et al., 2019; Savage, 1954; Tsetsos et al., 2016; Tullock, 1964; Tversky, 1969). Transitivity states that for any three alternatives, A, B, and C, if one prefers A to B, and B to C, then he or she must prefer A to C to be transitive. Transitivity of preferences seems so intuitive and obvious that any violation of it could be regarded as being counterintuitive or even irrational. Savage (1954) mentioned in his book on intransitivity that,

> "... when it is explicitly brought to my attention that I have shown a preference for f as compared with g, for g as compared with h, and for h, as compared with f, I feel uncomfortable in much the same way that I do when it is brought to my attention that some of my beliefs are logically contradictory (Savage, 1954, p. 21)."

In the real world, however, people wouldn't think much about transitivity of choices, or rationality at every moment when they make choices. When choices are trivial, people tend not to spend too much resources, but to rely on a simple decision rule to make decisions. When choices are complicated, people would find a way to produce satisfying outcomes without investing too much resources. Gigerenzer and his colleagues have studied such patterns of decision making, in which people sacrifice the quality of outcomes for the sake of effort reduction (Brandstätter, Gigerenzer, & Hertwig, 2006; Gigerenzer & Brighton, 2009; Gigerenzer & Goldstein, 1996; Gigerenzer & Selten, 2001; Gigerenzer & Todd, 1999). They called those decision-making strategies *heuristics*, which is characterized by the *less-is-more* framework (Gigerenzer & Brighton, 2009), or the effort-reduction framework (Shah & Oppenheimer, 2008). This style of making decisions is clearly contrasted with transitivity. Transitivity encourages decision makers to consider all relevant information, because there exists a possibility that missed out information might end up causing intransitivity of preferences. Heuristics, on the other hand, allow decision makers not to process all pieces of information, yet still allowing them to arrive at a satisfying decision. Some of such heuristic approaches have been generalized and axiomatized under the name "Lexicographic semiorder" (Davis-Stober, 2012; Luce, 1956; Tversky, 1969), where decision makers utilize attribute-wise comparisons between alternatives. In this chapter, I discuss transitivity and lexicographic semiorder as a means to provide a better understanding of rational theories and heuristic-based theories, respectively. Since transitivity and lexicographic semiorder describe clearly different decision-making processes, the present chapter aims to examine each theory to a great extent to help to grasp the essence of each theory.

In doing so, I largely follow Luce and Narens's insight about the axiomatic approach, where they stress the importance of testing axiomatic representational theories against empirical data. One of the challenges we face immediately when testing axioms against empirical data is that axioms are deterministic, while empirical data are variable. As Tversky (1969) pointed out, choices are inherently stochastic. Even when decision makers are presented with the same problem repeatedly, they sometimes make different choices. Axioms, on the other hand, make deterministic statements, such as "if A is true, then B is true," and those statements allow no variability. Upon this observation, Luce and Narens (1994) presented the following challenge as one of the 15 problems of measurement theory:

> PROBLEM 2. Specify a probabilistic version of measurement theory and the related statistical methods for evaluating whether or not a data set supports or refutes specific measurement axioms (Luce & Narens, 1994, p. 227).

Luce and Narens' (1994) challenge is twofold: to recast axioms as a statistical model, and to build appropriate statistical methods to test the model. In this chapter, I survey previous approaches for either (or both) of the two aspects of the problem. But, before I proceed, I would like to clarify a few things that may cause confusions or misunderstandings on the topic.

First, I will intentionally distinguish the use of the term "preference" from the term "choice" (or "decision," which I will use interchangeably with the term "choice"). In decision theory, preferences are used as states of mind that give rise to choices (Davis-Stober, Brown, & Cavagnaro, 2015). That is, preference is considered a theoretical concept, which we cannot observe directly, while choices are considered realizations of preference in the real world, which we can observe. This distinction is important, because when we deal with choice variability, which I will discuss in detail later in this dissertation, whether we locate the source of variability in preference or in choice directly determines what statistical model we should use for the given data.

Second, I exclusively consider choices under *risk* for the present study. In economics, there are three types of choices, according to the states of information available to decision makers; those are choices under *certainty*, choices under *ignorance*, and choices under *risk*. Choices under certainty refer to the choices, where decision makers are quite certain that their choices will result in specific outcomes. An example of choices under certainty is choose a menu for dinner at a restaurant. When you choose a certain menu, you are very certain that you will receive what you choose, so whether the outcome you expect will happen is not your concern with this type of choices. Choices under ignorance refer to the choices, where decision makers have no idea how likely their choices will bring about certain outcomes, and so they are not able to assign probabilities to those outcomes. An example of this type of choices is choose a life partner. When you ask someone out, you have no idea how likely you will be with this person for the rest of your life. And finally, choices under risk refer to the choices, where decision makers are able to assign probabilities to all the outcomes of each choice. Unlike the other two types of choices, choices under risk involve probabilities. That means, first, we can apply a vast literature on probability theory to our decision-making processes; second, studying this type of choices can provide a deep insight into our day-to-day lives because our daily lives are full of probability judgments (e.g., weather or economic forecasts). Note that probabilities here need not be mathematically rigorous (Resnik, 1987). That is, we would still be able to apply probability theory to our subjective probabilities about certain events.

Now I give a brief review of transitivity and lexicographic semiorders.

2.1 Transitivity and lexicographic semiorders

2.1.1 Transitivity and rationality

Transitivity is one of the fundamental axioms, upon which many decision theories rely. Specifically, it states that for any three alternatives, A, B, and C, if one prefers A to B, and B to C, then he or she must prefer A to C to be transitive. To better understand its necessity for decision theories, it helps to think why we choose one alternative over the other in the first place; that is, what makes one stand out among candidate alternatives. This is one of the problems economists have long been attempting to tackle (Irwin, 1958; Luce & Suppes, 1965). To this question, economists have developed a theoretical concept, utility, and assumed that decision makers would make a decision in a way that maximizes their utility. In other words, individual decision makers would compute utility for each available alternative and compare the computed value of utility with one another and choose the one with the highest value. Utility, in this sense, represents the desirability of the given alternative, often understood as self-interest (Mcfadden, 2001; Simon, 1978). Jeremy Bentham (1789), who first suggested the use of utility when reasoning human choices, specifically wrote about utility as follows,

> By utility is meant that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness (all this in the present case comes to the same thing), or (what comes again to the same thing) to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered: if that party be the community in general, then the happiness of the community; if a particular individual, then the happiness of that individual. (Jeremy Bentham, 1789, as quoted in Luce & Suppes, 1965)

This brilliant idea of utility enables economists to consolidate all different kinds of reasons as to why an individual prefers one alternative over the other into one single quantity, which has made the study of preferences quantifiable. Traditionally, economists have assumed a utility function that generates this single quantity from multiple reasons (May, 1954). This can be formally stated as follows; if we let $u(\cdot)$ be a utility function, which maps utility of an alternative onto a real number, then

$$A \succ B \text{ if and only if } u(A) > u(B). \tag{2.1}$$

One obvious assumption implied by such utility function is that we can assign a numerical value to each alternative according to our preferences. Under this system, it is quite easy to see how transitivity of preferences holds naturally, simply because numerical values are ordered in a transitive manner. Once an individual successfully assigns a numerical value (i.e., utility) to each of a set of alternatives, aligning them from the smallest to the largest according to their assigned values will immediately result, so will hold transitivity. Indeed, transitivity is a necessary condition for the existence of a utility function (Luce & Suppes, 1965; Tversky, 1969; Regenwetter, Dana, & Davis-Stober, 2011), and even sufficient in most practical situations, where the number of alternatives are finite (May, 1954). In other words, wherever transitivity does not hold, no utility function that satisfies Relation (2.1) exists, which effectively means that empirical tests of transitivity can serve as tests as to whether or not to justify the use of utility.

Despite its fundamental properties in utility-centered theories, transitivity has often been overlooked in empirical settings and taken for granted when theories are tested against empirical data. For an ideally rational agent, of course, transitivity of preferences will never be questioned as the rational agent would have no problem ordering alternatives according to utility and behaving accordingly. However, such rationality isn't always present at the moment of making decisions in empirical settings (Todd & Gigerenzer, 2000). One main factor that keeps the agent from forming transitive preferences is conflicting criteria (Shepard, 1964). Many choices in the real world often require the agent to consider multiple reasons with conflicting criteria. Buying a house in a good neighborhood costs much more than buying one in a not-so-good neighborhood; healthy foods in general taste bland compared to junk foods; big screen laptops hinder portability. When deciding between alternatives with conflicting criteria, the agent must sacrifice one reason for other reasons that he or she regards more important. In other words, he or she must make trade-offs between reasons. Making trade-offs between multiple reasons is a demanding task especially when the reasons are negatively correlated (Shanteau & Thomas, 2000). In such environments, seeking a good reason necessarily causes other reasons to be bad; for instance, a good neighborhood will make the price of a house increase. That is, no alternative can maximize all the reasons simultaneously (McClelland, 1978). If the agent struggles with deciding what to prioritize, he or she might apply one of the conflicting criteria this time and other one next time; that is, he or she may keep on switching between criteria haphazardly whenever the problem shows up, which potentially poses the agent the risk of forming circular patterns of preferences, or intransitivity of preferences.

For such reasons, transitivity is often considered a quality ascribed to rational decision makers. Intuitively, for anyone who prefers A to B and B to C, it is rational to predict that he or she would prefer A over C, not the other way around. If one is alleged to prefer A to B, B to C, and C to A, we can have him or her confronted with the three alternatives A, B, C all at once, and ask which one he or she prefers the most. Any answer will contradict his or her previous binary preferences, clearly not a feature we'd expect to see from rational decision makers (Anand, 1993; Savage, 1954; Tullock, 1964), or a desirable property of a decision rule (Tversky, 1969). Indeed, Tversky (1969) interviewed intransitive participants at the end of the experiment session, and they were embarrassed when Tversky had them face the intransitivity of preferences they had formed during the experiment. One participant attributed his intransitive preferences to a lack of attention or a mistake, denying the possibility that he had actually had intransitive preferences when responding to the task. There is also evidence that when people are confronted with their intransitive choices, they tend to modify their choices in accordance with transitivity (Luce & Raiffa, 1989).

Although transitivity is undoubtedly a cornerstone of rational choices, transitivity doesn't always provide the most effective principle for every situation. For example, when it comes to complex problems with lots of uncertainties, such as job offers, it is extremely difficult to make one's preferences transitive over all available alternatives (Tversky, 1969). Under such environments, we have no choices but to rely on uncertain inferences due to unknown factors (Todd & Gigerenzer, 2000), which may force us end up with intransitive choices. On the contrary, when the task is too trivial, such as deciding what to eat for lunch, satisfying transitivity by arranging all available alternatives from least favorite to most favorite could be an overkill. In those incidents, we'd just need to decide what to eat quickly, paying attention only to a few alternatives available at the moment. No one would question the agents' rationality because of the way they decide their lunch menu. Like these examples, our decision making processes are heavily affected by the characteristics of the environments we operate in (Simon, 1956). Depending on the problem we are trying to solve, we may possess only a limited amount of information with lots of unknowns, perhaps no optimal strategies exist or are needed in some cases. In this sense, Payne, Bettman, and Johnson (1993) argued that an intelligent decision maker would adapt his or her decision strategy specifically to the type of the task at hand, which sometimes encourages the use of a heuristic.

In the next section, I will review one of heuristic-based decision making strategies, *lexicographic semiorder*. Tversky (1969) proposed the theory based on the idea of semiorder (Luce, 1956) and used it to account for a pattern of transitivity violation that he observed during the experiment. The main motivation for the use of lexicographic semiorders was simply the fact that the lexicographic semiorder theory is able to account for a systematic violation of transitivity, but it's become one of the main heuristic strategies for complex tasks. In contrast to transitivity, the lexicographic semiorder theory provides approaches that allow decision makers not to process all pieces of information when making decisions, yet it still allows them to arrive at satisfying decisions. The lexicographic semiorder theory is particularly effective when the task is complicated with multi-attribute alternatives. I give a brief overview of the theory in the next section.

2.1.2 Lexicographic Semiorders and Heuristics

When the task requires agents to consider a vast amount of information, or when some attributes in the task are noisy or uncertain, agents may find it hard to make decisions the way most traditional decision theories (e.g., expected utility theory, von Neumann & Morgenstern, 1947) would state. Such decision theories in general propose utilizing all the available information, or cues, to make decisions. In the aforementioned cases, however, agents simply cannot consider all the information either because the information is too much to process, or because the information is too noisy to process. Instead, agents would employ simple rules, such as lexicographic semiorders (Tversky, 1969), in order to make decisions when having to deal with complex multidimensional alternatives. In this section, I discuss the lexicographic semiorder as a good approximation of one's true preferences when the environments don't allow the use of more rigorous decision theories. In doing so, I review why and how heuristic decision models like lexicographic semiorders work under such environments, and then I conclude the section with comparing heuristic decision strategies, which don't necessarily satisfy transitivity, to traditional decision theories, which require transitivity as one of its building blocks.

Lexicographic semiorder is a variant of lexicographic heuristic (Fishburn, 1974), to which Tversky (1969) added the system of semiorder (Luce, 1956). The way of the application resembles the way we'd find a word from a dictionary, so is the name "lexicographic." Lexicographic semiorder features a component-wise comparison, applying the semiorder principle to one attribute at a time, with all the attributes ordered according to agents' priority. If the attribute under consideration doesn't produce a decisive winner, the agent would move on to the next attribute in line and repeat the same process until one of the attributes favors one alternative over the other by more than the agent think indifferent. Thus, the process is also referred to as a non-compensatory strategy (Payne, Bettman, & Johnson, 1992) since a good value on one attribute cannot make up for a bad value on other attributes.

While the lexicographic part is responsible for how the theory explores multiple attributes, the semiorder part governs the way the attributes are compared between alternatives. Semiorder is a measurement system Luce (1956) developed after he brought his attention to Armstrong's (1950) remarks on intransitivity of indifference. Luce admitted that indifference hardly satisfies transitivity if the indifference is caused by our inability to perceive a small change. To illustrate this point, suppose we have multiple cups of black coffee lined up on the table and plan to add different amounts of sugar to each cup. If we add to the first cup a small amount of sugar, so small that our sense of taste can't biologically distinguish, then we'd be indifferent between the sugar-added coffee and black coffee. However, if we keep on increasing the amount of sugar by the same amount for the next cup after another, there will be a point where we can taste the difference, that is, no more indifference persists. In other words, the indifference between adjacent cups doesn't guarantee the indifference between the first and the last cups, implying intransitivity of indifference. Semiorder takes such intransitive interrelations as its core foundation in the context of utility discrimination. Humans are not sensitive to all changes in utility; only the changes greater than the limit our sense can tell affect our preferences.

Combining these two ideas, lexicographic semiorders are able to account for numerous choice patterns by varying two variables: order and threshold. First, the order variable concerns the order in which the attributes are considered by agents. Which attribute first comes into consideration highly depends on the agents' priority. Even for the same task, different individuals may take different orders in considering attributes and could end up with completely opposite answers. Second, the threshold variable determines the limit to which people think the alternatives are indifferent. In other words, any difference made under this limit will be obscure to the decision maker; only the difference greater than the limit will be significant to the decision maker, leading to a decision. As the theory concerns itself with one attribute at a time on the way to a final decision, the threshold variable needs a value to be set for every attribute in the alternative. If the decision maker finds no attributes making any difference that meets the threshold set by him or her, he or she will end up being indifferent between the alternatives under comparison. Based on these two variables, Regenwetter, Dana, Davis-Stober, and Guo (2011) formalized lexicographic semiorders and found that the theory is able to account for 111 distinctive choice patterns when there are five choice alternatives with two attributes to be considered.

The original purpose of Tversky's developing the theory, however, was not that the lexicographic semiorder is able to account for various decision patterns, but that the theory can provide an account of a systematic violation of transitivity. Under a certain context, Tversky (1969) found a few participants violate the principle of transitivity in a similar way. When faced with a task that has noisy information, these participants would ignore the noisy part of information and rely solely on the other aspect of information that is less noisy. For example, in the experiment Tversky conducted, he intentionally made the information regarding probability of winning of the gambles hard to distinguish one from another, or noisy; he represented it in a form of pie chart. He found that some participants ignored a small difference in probability of winning between gamble stimuli, because it is hard to tell such a small difference in the region of a pie. Instead, they based their decisions solely on the payoff information, which was written in numeric value, easily distinguishable. When the difference in the probability of winning grew larger, however, participants began to consider probability of winning for their choices, since the gamble stimuli were designed to have increasing expected values with the probability of winning. However, this particular pattern of choices can give rise to intransitive preferences.

Consider the gamble stimuli Tversky used in his experiments. Tversky made up this particular set of gambles in an attempt to induce intransitive patterns of choices.

The gamble has two attributes to consider: probability of winning and payoff (see Figure 2.1). Each gamble consists of different values of the two attributes, and the assigned values are negatively correlated with each other (i.e., no dominant alternatives exist), with its expected value designed to increase with probability of winning. The experiment was a series of binary choices tasks, where in each task, participants were asked to choose one between a pair of gambles, randomly selected from the gambles in Figure 2.1. To better understand how lexicographic smiorder leads to intransitive preferences in this particular setting, suppose a participant who makes choices according to the lexicographic semiorder principle. Since the probability of winning is represented in a pie format, it is not easy to tell the difference between the two given gambles, especially when the two gambles are adjacent. In fact, the difference in probability of winning between any adjacent gambles in Figure 2.1 is just 1/24. On the other hand, payoffs of the gambles appear in numeric value on top of each gamble pie, which provides easy comparison. And that's the type of information people would primarily consider when making choices, according to Tversky's (1969) account. That is, people would tend to rely on more distinguishable information, such as payoff in this example, when choosing between two competing alternatives, rather than ambiguous information, such as probability of winning. Following this account yields the following preferences between adjacent gambles: $A \succ B$, $B \succ C$, $C \succ D$, and $D \succ E$. However, when it comes to comparison between far-away gambles, say A and E, the participant would choose the one with higher probability of winning, i.e., $E \succ A$. This is because now the difference in probability is big enough to draw attention to it and the expected value greatly favors Gamble E. In other words, the participant has formed an intransitive preferences, $A \succ B$, $B \succ C$, $C \succ D$, $D \succ E$, and $E \succ A$.

Although it could generate intransitive preferences, lexicographic semiorder has its attractiveness to decision makers for the following reasons. First, it avoids making

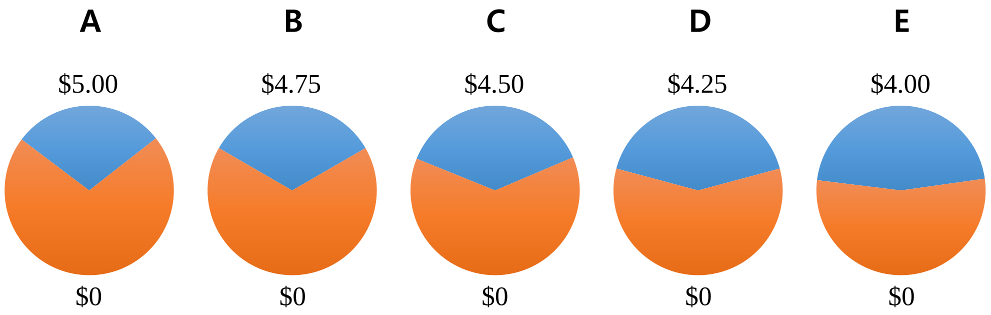


Figure 2.1: Gamble set employed in Tversky's (1969) experiment. Gambles are in a pie format, where the blue-colored area represents the probability of winning the specified payoff. The expected values of the gambles increase with the probability of winning.

trade-offs between attributes, which helps to reduce mental effort and cognitive burden (Shah & Oppenheimer, 2008; Tversky, Sattath, & Slovic, 1988). When offered with complex multi-dimensional alternatives, decision makers could be overwhelmed by the sheer amount of information to process and all the conflicts to resolve between multiple attributes. In this case, lexicographic semiorders can help because it allows decision makers to consider one attribute at a time, even with the possibility that the first few cues could produce a final decision (Shah & Oppenheimer, 2008).

Second, lexicographic semiorder is easy to justify or apply in the real world. This is largely due to the fact that the lexicographic semiorder is considered a *procedural theory* (Starmer, 2000), which, in contrast to conventional theories in economics, concerns itself with the actual processes that are used to make choices. Theories of this class provide decision makers with precise directions to follow in order to make choices. Lexicographic semiorders, for example, state what to do with respect to the task at hand; such as, search for a dominant alternative if any, or look for the alternative that is superior on the more important attribute otherwise (Tversky et al., 1988). Having a set of clear rules in terms of making choices can be a great appeal to decision makers over conventional strategies especially when the task is complex

(Payne et al., 1993; Tversky, 1969).

Lastly, lexicographic semiorders in general take less time to reach a choice. As mentioned above, lexicographic semiorder circumvents the need for making trade-offs between conflicting attributes. In addition, there is even a possibility that decision makers could arrive at a decision with just a few pieces of information. This helps decision makers not just reduce the cognitive burden regarding the task, but also make decisions fast. Rieskamp and Hoffrage (2008), for example, found that under the high time-pressure condition, participants tended to use simple heuristics, such as lexicographic semiorder, in their experiments. They reasoned that it is simply not possible for decision makers to use a strategy that utilizes all available pieces of information, when there is not enough time. In that case, decision makers would instead turn to a strategy that helps them arrive at conclusions based upon only a few cues, or without having to make trade-offs.

The above benefits can be discussed within the effort-reduction framework proposed by Shah and Oppenheimer (2008). Most decision theories in economics assume an agent who possesses unbounded rationality with unlimited resources and unlimited time. Such an ideally rational decision maker would use a complex algorithm to arrive at an optimal decision, where the complex algorithm here usually refers to the weighted additive model (Brandstätter et al., 2006; Keeney & Raiffa, 1993; Payne et al., 1993). According to the weighted additive model, decision makers first examine all of the available alternatives and related cues for each alternative. Then they weight each cue according to its contribution to the goal of decision. Lastly, decision makers integrate each cue's value multiplied by its weight to yield the overall value for an alternative. All these processes of the weighted additive model require great mental effort. As we humans in the real world are limited by our cognitive resources and the task environments, using such a complex algorithm as the weighted additive model is not always possible. Instead, we seek to reduce our effort associated with decision processes, which naturally brings our attention to the use of heuristics. In this sense, Shah and Oppenheimer argued that all heuristics involve some features of the effort-reduction principles. Lexicographic semiorders, for instance, involve methods of examining fewer cues and integrating less information. Although lexicographic semiorders may require all pieces of information to be examined before a decision is produced, it still allows decision makers to consider only one cue at a time, significantly reducing the amount of information to be processed at the time of comparison.

As we've discussed thus far, heuristics are mainly for helping decision makers arrive at satisfying decisions with less effort. As Simon noted earlier, humans do not possess unlimited rationality, but rather *bounded rationality* (Simon, 1955, 1990). Heuristics work because humans have limited cognitive resources, as well as limited time and information (Gigerenzer & Goldstein, 1996). People are simply not capable of applying complex rules to every task they face. Instead, they would ignore information when the tasks are complicated, which ironically often provides better solutions to the tasks (Gigerenzer & Brighton, 2009; Gigerenzer & Goldstein, 1996). Characterized by the less-is-more principle, Gigerenzer and Brighton (2009) argued that heuristics have helped humans survive in this complicated world, from an ecological perspective.

However, this very notion gives rise to the following fundamental questions: Do people actually use heuristics in the real world for their tasks? If so, does everyone use heuristics, or are there individual differences in the use of heuristics? These questions seem straightforward to answer, but testing whether people actually use heuristics is quite challenging. This is because heuristics, or any kind of decision theories, make deterministic statements while observed data are variable. Thus, testing decision theories must involve specifying statistical models for decision rules of interest and running an appropriate statistical test of the specified model on the observed data.

As mentioned above, these two matters have remained a challenge in the field of judgment and decision making (1994). In the rest of this chapter, we will examine how deterministic theories can be tested against empirical data. We first start with surveying different ways of interpreting the source of variability of the observed data, since how we see the variability is crucial in determining the statistical model for decision theories (Hey, 2005).

2.2 Statistical models for decision theories

Once Tversky (1969) observed, people are not consistent in their making decisions even when faced with the same task repeatedly. Those inconsistencies can occur without a systematic change in preference, but it could just be fluctuations in mind or inherent variability in choice (Loomes, 2005; Tversky, 1969). That is, people don't seem to want to make the same choice every time they are confronted with the same problem, or they'd just choose the wrong one by mistake. This is true even when people are put in a controlled experimental setting and asked to make repeated choices (Davis-Stober et al., 2015; Hey & Orme, 1994). It is such inconsistencies that make testing decision theories against empirical data challenging, because decision theories make deterministic statements. Then, one important question arises: How do we have to deal with the inconsistency inherent in data when testing decision theories? To answer this question, understanding the nature of variability of the data is the first step to take, as the story of variability unfolded by the data plays an important role in determining the statistical model for theories (Hey, 2005). And depending on the statistical model we use, our statistical inferences about the data will vary. In this section, we explore different ways to treat variability, or stochastic structure, of the data in the context of statistical modeling.

2.2.1 Variability as error

"Trembling hand" error

One of the most obvious ways to interpret the inconsistencies of empirical data is to treat it as error. From this point of view, decision makers are assumed to have a fixed "true" preference, but their choices are probabilistic, making occasional mistakes for various reasons. They might have been careless at the moment of making choices or misread the information about the given alternatives. All these carelessness or mistakes are assumed to lead to choices on the opposite side of the "true" preference, and hence considered error. A simple way to implement such error is via assigning a "trembling hand" type of error ε, to a fixed preference (Harless & Camerer, 1994; Loomes, 2005), where ε is a fixed probability, representing all different kinds of mistakes that can happen in the real world. Under this specification, it is the error parameter ε that accounts for all departures from the true preferences. In most cases, the value of ε is held constant across different pairs of alternatives, reflecting that those mistakes would happen at the same rate no matter the alternatives.

To illustrate how this specification of error works in a decision problem, suppose a choice between Gamble A (60% chance of winning \$30 and 40% chance of winning nothing) and Gamble B (80% chance of winning \$20 and 20% chance of winning nothing). If we use the expected utility theory (von Neumann & Morgenstern, 1947) for prediction, we need to compute each alternative's expected utility value by weighting respective utility values of payoffs by its probability and summing the weighted utility values. That is, the expected utility of Gamble A, $U(A)$, can be obtained by:

$$U(A) = \sum u(x_i^A)p_i^A,$$

where $u(\cdot)$ is a utility function that returns the utility of each payoff, x_i^A is the i-th payoff of Gamble A, and p_i^A is the probability that the i-th payoff of Gamble A is realized. Thus, the expected utility of Gamble A, $U(A)$, is given by $u(\$30).6 + u(\$0).4$, which would become $u(\$30).6$ if we assume no monetary outcome (i.e., \$0) produces zero utility. In the same way, the utility of Gamble B, $U(B)$, is given by $u(\$20).8$. Now the choice will be based on the comparison between $u(\$30).6$ and $u(\$20).8$ and which one to choose relies heavily upon what type of utility function $u(\cdot)$ we have assumed. There might be a number of factors that affect the shape of the utility function, and depending on what utility function we employ, the prediction may vary. But, once the utility function is decided, and the choice is predicted by the theory, the theory itself can't account for any deviations from its prediction. That is, whenever the same problem appears, the expected utility theory will predict the same choice for the same individual over and over again, with no exceptions allowed. In order for the model to account for potential inconsistencies of choice, we can assume the aforementioned fixed error rate on the predicted preferences. Suppose the theory predicts one's preference to be A over B, then we could assign a probability of error, or mistake, ε to that preference, so the probability of a decision maker choosing A over B, $P(A, B)$, becomes $1 - \varepsilon$. Equivalently, we can say that the probability of the same decision maker choosing B over A by mistake, $P(B, A)$, would be ε.

One main drawback of this *trembling hand* notion of error is that the error rate is fixed across all pairs of alternatives. No matter what pairs of alternatives appear, the same error rate applies according to this specification, predicting the same proportion of the choices to be made by mistake. It is not hard, however, to imagine situations where the error rate changes depending on what alternatives are being compared. For example, people tend to make more errors when they have to choose between similarly preferred objects than to choose between objects that differ greatly in preference. As Loomes (2005) argued, practicing one's preferences could be analogous to weightlifting in the sense that both are the judgments people make about the physical world. That is, when asked to guess which object is heavier between the two, people will likely make more errors when the two objects being considered weigh about the same than when one object weigh notably more than the other. The error rate, in this case, should reflect what pair of alternatives are given and vary accordingly, but the trembling hand error can't accommodate this need for flexibility. This is the primary reason why the trembling hand type of error isn't considered a viable way of the error specification for the principal model (Loomes, 2005; Loomes, Moffatt, & Sugden, 2002). But, its ease with which one can model the stochastic structure of empirical choices using just one parameter sometimes allows researchers to tackle more complicated data; Harless and Camerer (1994), for example, employed this trembling hand error for their meta-analysis because this particular error specification provided mathematical convenience when the authors attempted to analyze aggregated data over several studies.

Again, the trembling hand error is just one type of error specification with respect to the variability in empirical choices. Its simplicity can provide mathematical advantages in fitting theories of interest to data, but its assumption about the error rate being fixed, that is, independent of the uniqueness of each pair of alternatives, has raised skepticism about its plausibility in some contexts. In the next section, we discuss a different approach in modeling the stochastic component in data, which incorporates the feature of varying error rate depending on what pair of alternatives are being considered.

Error as "White Noise"

When Hey (2005) emphasized the importance of careful practice in specifying the nature of noise in data, he noted that the fixed error rate assumption wouldn't be able to account for those situations where the error rate should reflect the nature of the questions being considered. He argued that depending on the data we attempt to analyze with decision theories, we may need different stories about noise in data, which will bring about different implications of the results. The second approach in modeling noise is to add a “white noise” to the true preferences. On the surface, it sounds similar to the first approach, i.e., a constant error rate, but it assumes a slightly different structure as to modeling one’s choices. First, one’s true preference is determined by the valuations obtained by various preference functional forms; here the preference functional form comes from the formulation of various decision theories, so the resulting values will vary depending on which theory we employ to account for the data. Second, a normally distributed error with zero mean and variance of 1, called a white noise, is added to these computed preferences, reflecting that decision makers have chances of making mistakes when accessing to what they truly prefer. The observed choices for the given problem are then made based on the final values. Note that under this specification, decision theories serve to determine how to appraise one’s preferences via preference functional forms; that is, this part of the equation remains deterministic. All the variability in choices comes from the white noise part, and thus, the observed variability is assumed to follow the same distributional assumptions of the white noise.

To better understand how this approach works, consider a simple example, where a decision maker is confronted with a choice between A and B, as in the example mentioned above. If we let y be the value that determines the decision maker’s stated preferences, then y is given by:

$$y = V(A, B) + \varepsilon, \tag{2.2}$$

where $V(A, B)$ is the valuation of the chosen preference functional form for the choice between A and B, which is often regarded as the decision maker’s true preference, and ε is a white noise, distributed as $N(0, 1)$. At first glance, the valuation term $V(A, B)$ makes a major contribution to y since the white noise term only can make an impact on the stated preferences so much with its mean being fixed at 0 and its variance at 1.

In other words, if the two alternatives under consideration differ greatly in preference, there is little chance for the white noise to reverse one's true preference, resulting in a strict choice even when the same problem is asked repeatedly. Only when the two alternatives come close, so $V(A, B)$ is closer to 0, the white noise starts to affect the stated preferences, leading the observed choices to be more stochastic. If $V(A, B)$ is exactly equal to 0, then the choice, under this specification, will be random, where only the white noise plays a role in the choice process. This way, the white noise approach incorporates the feature of varying error rate depending on the alternatives being compared, providing a useful framework in modeling variability of choices for some decision tasks (Hey & Orme, 1994).

Although the white noise approach is considered one of the main ways of modeling noise in preferential choice data in economics (Loomes & Sugden, 1998), its origin dates back to Fechner (1860) in psychology, who endeavored to measure sensations with respect to the physical magnitude of stimuli, such as light, sound, weight and distance. In its simplest form, the Fechnerian model can be written as a probabilistic model that measures our ability to discern the magnitude between two physical stimuli. Suppose, for example, that one is asked to judge which is brighter between two stimuli that differ in light intensity. If the intensity of light A and light B is represented by real values v_A and v_B, respectively, then the probability $P(A, B)$ of the person choosing A over B is given by:

$$P(A, B) = \Phi(v_A - v_B), \tag{2.3}$$

where $\Phi(\cdot)$ is a cumulative function of the standard normal distribution. A close examination reveals that the Fechnerian model is mathematically equivalent to Equation 2.2, provided that the error ε in Equation 2.2 is distributed as $N(0, 1)$ for all pairs of the alternatives. This equivalence of the two models implies one easily overlooked point: forming preferences is also a type of judgments about the physical world around us, as with the judgments about light or sound (Loomes, 2005). That is, human judgments about preferences may well be noisy, especially when the two alternatives are measured about the same on the scale of preference. In that sense, the current white noise approach has its strength in that it provides the means to capturing such varying error rates.

While the white noise approach takes care of the stochastic structure of data, the problem of determining one's core preferences still remains. As mentioned above, this is where decision theories come into play. The valuation part of Equation 2.2, $V(A, B)$, is responsible for factoring in our preferences, which may vary depending on what theories we choose to apply. For illustration, let's consider the gamble example from the previous section again. Forming a preference between Gamble A ($30, .6; $0, .4) and Gamble B($20, .8; $0, .2) can be different from individual to individual depending on how one values the given monetary values, or how one views the probabilities associated with each outcome. It is decision theories' job to take into account all the related factors of one's forming preferences and produce the magnitude of preference, represented by a single number, for the given alternative. Then, we compute the net advantage of the pair of alternatives under consideration, which determines the value of $V(A, B)$. For example, the expected value, perhaps the simplest model of all, although it still serves as a good proxy for the expected utility (Hey, 2005), is given by the monetary value multiplied by its probability, which gives $18 for Gamble A and $16 for Gamble B. The simplest way to compute the net advantage is to take the difference of those obtained values for the given gambles, which results in the net advantage of $\$18 - \$16 = \$2$ in favor of Gamble A. Thus, decision makers would form a preference for Gamble A according to the model of our choice, although stated preferences will be subject to the aforementioned white noise.

The above example demonstrates the simplest calculation of $V(A, B)$, but the choice of the model is left to researchers. This is one benefit of the white noise approach in analyzing data, because it allows researchers to have such flexibility in choosing a core decision model to accommodate the goal of analysis. Taking advantage of this point, for example, Hey and Orme (1994) were able to compare the performance of 11 different decision theories in light of empirical data and to conclude that the expected utility plus error model stood out among other highly respected decision theories in the field. The authors suggested that it's the error model that should take credit for the expected utility model's success in accounting for the empirical data, because the pattern of the data hadn't seemed to support the expected utility theory until the white noise began to do its job capturing all random departures from the core preferences (see Loomes, 2005, for more detailed examples).

One problem of this approach, however, arises from the assumption that every decision maker has a core preference during a decision task, upon which all his or her choices based. It sounds natural to have a preference that we could rely on whenever we make a choice for a similar task, but sticking to one preference throughout a whole experimental session may not sound intuitive in some situations or at least for some individuals. For example, one may value the amount of payoff more than the probability of winning for some tasks, but the same individual may behave in the opposite way for some other tasks. In other words, some decision makers wouldn't make choices based only on a single preference, but they would instead employ a set of preferences, rather than one, and pick out one every time they are confronted with a new, but similar task. This way of decision making cannot be modeled by the true-preference-plus-error approach we've been discussing, since this approach assumes individuals to have only one preference throughout the given session. Also, if any variability occurs under the assumption of multiple preferences, we can't call it an error because they may have employed a different preference than before; that is, they just changed their mind. This is the exact pattern of what economists call "Random Preferences" (Loomes & Sugden, 1995). In the next section, we'll address this completely different way of accounting for choice variability.

2.2.2 Random preference

The random preference model (Loomes & Sugden, 1995) originates from the random utility model (Becker, Degroot, & Marschak, 1963), which accounts for choice variability via decision makers' varying preferences. According to the model, decision makers would have the alternatives aligned by preference, but such an ordering is considered random and subject to change every time they make choices. There might be various reasons why one has such varying preferences; it could be imprecise preferences that cause decision makers to change their preferences occasionally; or it could be task-related factors, such as the way the information about the alternatives are presented on screen, that encourage decision makers to apply a different preference to a different task. Although the interpretation as to why individuals would change their preferences may vary from study to study, those who employ the random preference model share the same perspective on inconsistent choices: the choice variability shouldn't be treated as error.

Suppose that we have a number of alternatives to choose from, $A, B, ...$, and that R is a set of all preference relations on those alternatives. Then the model assumes an additive probability measure Φ on R such that we measure the probability of any subsets r of R with Φ, i.e., $\Phi(r)$. When a decision maker is asked to make a choice between two alternatives, say, A and B, the model calls for all the preference relations from R that indicate the relation between A and B. Therefore, if the decision maker has chosen A over B, then the model computes the probability of the choice, $\Pr(A, B)$, by measuring probabilities of all the preference relations that imply $A \succ B$, i.e., $\Phi(A \succ B)$. If the decision maker chose B over A when the same task showed up again, the model would account for this choice by computing the probability of the preference relations that imply $B \succ A$. In other words, the model finds such varying preferences to be the source of inconsistent choices.

With varying preferences, testing the core decision theory of interest against datat comes down to testing all preference relations predicted by the theory. As we know, every decision theory predicts certain preference relations among the given alternatives. Expected utility theory, cumulative perspective theory, regret theory, rank-dependent utility theory, and so on, all theories describe different processes of decision making and contain different types of parameters in their models, but at the end of the day, they predict preference relations. More precisely, each decision theory restricts preference relations such that there exists a subset r of preference relations R, $r \in R$, where each preference relation in r conforms to the theory of interest. Therefore, testing the theory can be done via setting the area of the subset r to be 1 and running a statistical test to see how well the data can be generated by preferences from r. This effectively means that researchers vary values of every parameter of the theory of interest to see if the model can generate the data.

The downside of this approach, however, is there is a pattern of data a theory can't account for no matter how much researchers vary its parameters. For example, expected utility theory has only one parameter to vary, u, utility of a gain. Researchers could use various utility functions to account for the data. Moreover, they could even vary this parameter per individual, so that the expected utility theory can generate so many preference relations specifically adjusted for each individual's tendency toward decision making. But, if one of the gamble stimuli stochastically dominates the other one, that is, none of the values of the paramter predicts the opposite relation, the choice for this gamble pair will become deterministic, not stochastic. In this case, a single observation of the opposite choice can fail the expected utility theory, even when that choice is actually made by mistake. Acknowledging the drawback of the random preference model, researchers frequently combine the random preference model with tremble hand error model or Fechner model, in such cases that the random preference model alone cannot account for all the choice variability in data (Loomes et al., 2002; Loomes & Sugden, 1998).

2.2.3 Statistical model of axiomatic representational theories

Specifying the statistical model of a theory is a necessary and important step to take before one tests the theory against data. In the previous section, we've discussed three ways to deal with the stochastic component in data when testing core decision theories. Now we turn our attention to representational theories of preferences: transitivity and lexicographic semiorder. Transitivity is one of the axioms Von Neumann and Morgenstern (1947) used to develop the expected utility theory. Lexicographic semiorder (Tversky, 1969) is an attribute-wise decision making strategy that allows for intransitive preferences. Such representational theories state algebraic relations over the given alternatives, and so they provide no room for variability in data just like decision theories. Therefore, we need to formulate probabilistic versions of measurement axioms to be tested against data.

To this end, I introduce two approaches in this section: *error* model and *mixture* model. These models are similar in spirit to the ones from the previous section, the true-preference-plus-error model and the random preference model, respectively; but, they are different in that the error model and the mixture model of the representational theories are often constructed via collections of linear inequality constraints over binary choice probabilities (Davis-Stober, 2009; Zwilling et al., 2019). Thus, when we test those theories, we need to test sets of inequalities of choice probabilities. And that requires us first to define representational theories of interest in terms of a set of appropriate inequalities. In this section, I discuss the error model and mixture model of transitivity and lexicographic semiorders.

Error model

Error model of transitivity One of the simplest statistical models of transitivity comes from the assumption that choices people make are variable. In this approach, the model doesn't seek to answer why people's choices vary; instead, it focuses on the very fact that people's choices vary, that is, the same decision maker makes different choices even when he or she is confronted with the same task repeatedly. Based on this observation, Tversky (1969) made a statement about inconsistent choices:

> "... the observed inconsistencies reflect inherent variability or momentary fluctuation in the evaluative process. This consideration suggests that preference should be defined in a probabilistic fashion (p. 31)."

He then defined preference in a probabilistic way:

$$A \succ B \text{ if and only if } P(A,B) > 0.5. \tag{2.4}$$

In other words, if one is to say A is preferred to B, then he or she must choose A more than half the time when the pair of A and B is given multiple times. Then, we can re-state transitivity in terms of the above probabilistic version of preference: for any triplet of choice alternatives A, B, and C,

$$\text{If } P(A,B) \geq 0.5 \text{ and } P(B,C) \geq 0.5, \text{ then } P(A,C) \geq 0.5.$$

This is the most general probabilistic version of transitivity, called Weak Stochastic Transitivity (WST; Block & Marschak, 1960). There are more restricted versions of stochastic transitivity as well, namely Moderate Stochastic Transitivity (MST), and Strong Stochastic Transitivity (SST).

$$\text{MST: } P(A,B) \geq 0.5 \text{ and } P(B,C) \geq 0.5 \Rightarrow P(A,C) \geq \min\{P(A,B),\ P(B,C)\},$$

$$\text{SST: } P(A,B) \geq 0.5 \text{ and } P(B,C) \geq 0.5 \Rightarrow P(A,C) \geq \max\{P(A,B),\ P(B,C)\}.$$

As the names suggested, these models place more restrictions over the space of choice probabilities, making it easy for data to reject the model. These stronger versions of stochastic transitivity can provide tougher standards for deciding whether or not the data are transitive, but in this section, I exclusively consider WST for our error model of transitivity for the following reasons. First, WST is one of the most widely researched transitivity model; whenever researchers are interested in testing transitivity of data, WST is the one that most of them would turn to (e.g., Mellers & Biagini, 1994; Myung, Karabatsos, & Iverson, 2005; Tsetsos et al., 2016). Also, WST is defined in a straightforward and intuitive fashion (i.e., the binary preference relations in the original transitivity axiom are simply replaced by Equation (2.4)), which provides straightforward interpretation, as well as mathematical ease with testing the theory. Due to its structure of the definition, the WST is cast as a special case of the "supermajority specification" under the QTest framework (Regenwetter et al., 2014; Zwilling et al., 2019), which I employ as the main analysis tool in later chapters.

Although we now have a statistical model of transitivity, that doesn't automatically lead to a set of inequalities readily testable against data. In order to achieve this, we need more information about the setting where the data are collected. For example, transitivity is a relation defined over triplets of alternatives. So, how many triplets we can make out of the given alternatives is a critical factor for testing transitivity against data. If there are three alternatives A, B, and C, the number of transitivity relations over the set $\{A,\ B,\ C\}$ equals $3! = 6$. If there are five alternatives A, B, C, D, and E, the number of transitivity relations over the set $\{A,\ B,\ C,\ D,\ E\}$ equals

$\binom{5}{3} \times 3! = 60$. The number of transitivity relations increase by a factor of the number of alternatives under consideration, so including a large number of alternatives could present a challenge for testing transitivity against data, even with the simplest model of transitivity.

For an illustration purpose, suppose a set of three alternatives $\{A,\ B,\ C\}$. Transitivity holds if a decision maker has one of the following preferences: $A \succ B \succ C$ (abbreviated to ABC), ACB, BAC, BCA, CAB, and CBA. Note that this list of preferences is simply the permutation of the three given alternatives. Choices are variable, so even if the decision maker has a preference of ABC, he or she may choose B when A and B are offered. Thus, we employ the WST to test transitivity against empirical data. But, we are not certain which preference ordering the decision maker would have in mind. Thus, we need to consider WSTs in terms of all different preference orderings, which reveals the full complexity of the model we aim to test (I write P_{AB} for $P(A, B)$ for brevity):

$$\begin{aligned}
\Big[&\{(\mathrm{P}_{AB} > 0.5) \text{ and } (\mathrm{P}_{BC} > 0.5) \text{ and } (\mathrm{P}_{AC} > 0.5)\} \text{ or} \\
&\{(\mathrm{P}_{AC} > 0.5) \text{ and } (\mathrm{P}_{CB} > 0.5) \text{ and } (\mathrm{P}_{AB} > 0.5)\} \text{ or} \\
&\{(\mathrm{P}_{BA} > 0.5) \text{ and } (\mathrm{P}_{AC} > 0.5) \text{ and } (\mathrm{P}_{BC} > 0.5)\} \text{ or} \\
&\{(\mathrm{P}_{BC} > 0.5) \text{ and } (\mathrm{P}_{CA} > 0.5) \text{ and } (\mathrm{P}_{BA} > 0.5)\} \text{ or} \\
&\{(\mathrm{P}_{CA} > 0.5) \text{ and } (\mathrm{P}_{AB} > 0.5) \text{ and } (\mathrm{P}_{CB} > 0.5)\} \text{ or} \\
&\{(\mathrm{P}_{CB} > 0.5) \text{ and } (\mathrm{P}_{BA} > 0.5) \text{ and } (\mathrm{P}_{CA} > 0.5)\}\Big].
\end{aligned}$$

This model is exactly what Regenwetter and his colleagues (2010) spelled out for the WST model when the number of alternatives equals three. As mentioned above, the number of transitivity relations increases with the number of alternatives considered. That is, when the number of alternatives is 5, the number of transitivity relations becomes 60, which provides 60 different inequalities to be tested against data. This is undoubtedly not an easy task. But, recently, there have been many approaches developed to tackle this kind of problems (e.g., Davis-Stober, 2009; Heck & Davis-Stober, 2019; Myung et al., 2005; Silvapulle & Sen, 2005). Among them, the QTest (Regenwetter et al., 2014; Zwilling et al., 2019), a public domain software package, provides a general modeling framework for probabilistic binary choices. Testing collections of multiple inequalities against data is where the QTest excels, so I exploit the QTest framework to analyze choice data in later chapters.

Error model of lexicographic semiorders In this section, I discuss the error model of lexicographic semiorders. The same modeling principle applies to lexicographic semiorder, as, like transitivity, lexicographic semiorder only makes deterministic predictions. Particularly, lexicographic semiorder uses decision makers' priorities among the given attributes and thresholds of the attributes to predict preference relations, but lexicographic semiorder alone doesn't account for variability of data. Thus, we apply to the lexicographic semiorder the same error model as we did for transitivity, which yields the Lexicographic Semiorder Error Model (LSEM; Davis-Stober et al., 2019; Park, Davis-Stober, Snyder, Messner, & Regenwetter, 2019). The LSEM allows decision makers to make any mistakes in executing preferences through choices up to the pre-specified error rate. Suppose a set C of alternatives $\mathcal{C} = \{A,\ B,\ C, ...\}$. Then, we write LSEM as follows: for any A and B in Set $\mathcal{C}$,

$$A \succ_{LS} B \text{ if and only if } P(A, B) > 1 - \lambda, \tag{2.5}$$

where $\succ_{LS}$ is a binary preference relation predicted by lexicographic semiorder, and λ is a pre-specified error rate. When λ is set to 0.5, the above definition will become much similar to Equation (2.4), and so the LSEM is frequently considered a lexicographic semiorder's counterpart to WST for transitivity (Davis-Stober et al., 2019; Park et al., 2019).

As with transitivity, deriving lexicographic semiorder relations relies on the specific setting where the data are collected. This is because the lexicographic semiorders utilize the attributes of alternatives to predict preferences, and its predictions will be different if different alternatives are employed. Consider, for example, a set of three gambles: Gamble A (\$6.00, .29), Gamble B (\$5.00, .35), Gamble C (\$4.00, .41). Suppose that a decision maker prioritizes probability of winning over payoff, and that her threshold for recognizing the difference in probability is 0.1. That is, any difference in probability smaller than 0.1 is negligible to her, which makes her indifferent between any pair of probabilities that differ by less than 0.1. Also, suppose that her threshold for payoff is \$0.50. In this case, when the gamble pair A and B is offered, she would prefer A to B, because Gamble A's probability of winning differs from Gamble B's by less than her threshold, 0.1, but A offers better payoff than B, with the difference well exceeding her threshold for payoff, \$0.50. Applying the same process to the rest of the gamble pairs, we could obtain the following preferences: $B \succ C$ and $C \succ A$.

The above example demonstrates how we obtain lexicographic semiorders over the set $\{A,\ B,\ C\}$ for a particular decision maker who prioritizes probability of winning and has thresholds of 0.1 for probability of winning and \$0.50 for payoff. By varying the order of priority and threshold for each attribute, we could obtain the complete list of lexicographic semiorders over the set $\{A,\ B,\ C\}$, which are shown in Table 2.1.

The first three columns in Table 2.1 illustrate all possible cases that different levels of priority and thresholds can generate in the given setting. In this setting, the choice alternatives are monetary gambles, where probability of winning and payoff are the only pieces of information available. So, when it comes to priority, one simply

Table 2.1: Full list of lexicographic semiorders over $\{A,\ B,\ C\}$

	Thresholds (τ)		Pairs of alternatives		
Priority	Probability (τ_1)	Payoff (τ_2)	A vs B	B vs C	A vs C
Probability of winning	$0 \leq \tau_1 \leq 0.06$	$\$0.00 \leq \tau_2 \leq \$1.00$	B $\succ$ A	C $\succ$ B	C $\succ$ A
		$\$1.00 < \tau_2 \leq \$2.00$	B $\succ$ A	C $\succ$ B	C $\succ$ A
		$\$2.00 < \tau_2$	B $\succ$ A	C $\succ$ B	C $\succ$ A
	$0.06 < \tau_1 \leq 0.12$	$\$0.00 \leq \tau_2 \leq \$1.00$	A $\succ$ B	B $\succ$ C	C $\succ$ A
		$\$1.00 < \tau_2 \leq \$2.00$	A $\sim$ B	B $\sim$ C	C $\succ$ A
		$\$2.00 < \tau_2$	A $\sim$ B	B $\sim$ C	C $\succ$ A
	$0.12 < \tau_1$	$\$0.00 \leq \tau_2 \leq \$1.00$	A $\succ$ B	B $\succ$ C	A $\succ$ C
		$\$1.00 < \tau_2 \leq \$2.00$	A $\sim$ B	B $\sim$ C	A $\succ$ C
		$\$2.00 < \tau_2$	A $\sim$ B	B $\sim$ C	A $\sim$ C
Payoff	$0 \leq \tau_1 \leq 0.06$	$\$0.00 \leq \tau_2 \leq \$1.00$	A $\succ$ B	B $\succ$ C	A $\succ$ C
		$\$1.00 < \tau_2 \leq \$2.00$	B $\succ$ A	C $\succ$ B	A $\succ$ C
		$\$2.00 < \tau_2$	B $\succ$ A	C $\succ$ B	C $\succ$ A
	$0.06 < \tau_1 \leq 0.12$	$\$0.00 \leq \tau_2 \leq \$1.00$	A $\succ$ B	B $\succ$ C	A $\succ$ C
		$\$1.00 < \tau_2 \leq \$2.00$	A $\sim$ B	B $\sim$ C	A $\succ$ C
		$\$2.00 < \tau_2$	A $\sim$ B	B $\sim$ C	C $\succ$ A
	$0.12 < \tau_1$	$\$0.00 \leq \tau_2 \leq \$1.00$	A $\succ$ B	B $\succ$ C	A $\succ$ C
		$\$1.00 < \tau_2 \leq \$2.00$	A $\sim$ B	B $\sim$ C	A $\succ$ C
		$\$2.00 < \tau_2$	A $\sim$ B	B $\sim$ C	A $\sim$ C

Note. Lexicographic semiorder exploits priority among attributes and thresholds of the attributes to form preferences. First three columns show all possible combinations of priority and thresholds of the given attributes in the particular setting described in text, where each row corresponds to a unique combination of priority and thresholds. The last three columns show binary preference relations predicted by lexicographic semiorders, according to the combination of priority and thresholds.

decides which of the two attributes (i.e., probability of winning and payoff) is of more importance, which is shown in the first column. For thresholds, only the three ranges of thresholds are valid for each attribute here due to the number of alternatives under consideration. First, one recognizes every difference between any gamble pairs, even for the adjacent gamble pairs, A and B, and B and C, which provides the smallest threshold of each attribute. Second, one may be indifferent between adjacent gamble pairs, but the difference between A and C is big enough to produce a preference.

Lastly, one may be indifferent between any gamble pairs, because the differences are not big enough. Based on the combinations of priority and thresholds, the last three columns show the preferences for each gamble pair predicted by lexicographic semiorders. Note that although every row has different combinations of priority and thresholds, some of the combinations lead to the same preferences. This is because lexicographic semiorder is a non-compensatory decision-making strategy, meaning that if the prioritized attribute is decisive, other attributes can't play a role in making a decision no matter how big of a difference they make. For example, if a decision maker prioritizes probability of winning and finds one gamble favorable in that aspect, then any difference in payoff cannot reverse this decision regardless of which tresholds for payoff the decision maker may have. Hence, the first three rows in the last three columns all have the same preferences.

For a statistical model of lexicographic semiorders, we simply replace the preference relations in Table 2.1 with the probabilistic preference defined in Equation (2.5). Then, each row of Table 2.1 serves as an inequality of LSEM. As with transitivity, the LSEM gets complicated as the number of alternatives under consideration increases. Hence, testing the LSEM against data specifically when the number of alternatives grows larger presents a non-trivial problem. In the statistical inference section, I review the methods of testing those inequalities against data in either the frequentist way or Bayesian way. But, before we proceed, I would like to discuss a different statistical model, with a completely different assumption about the choice variability: the mixture model. The mixture model, like the error model, consists of a set of inequalities that put constraints over the space of binary choice probabilities. But, the way it sets up the inequalities is different, and the difference stems from the different assumption as to why choices vary. In the next section, I examine the mixture model of transitivity and lexicographic semiorders.

Mixture model

While the error model attributes variable choices to error, the mixture model attributes variable choices to one's having multiple preferences. According to the mixture model, individuals are not assumed to make errors, yet their choices still vary because they are "*in different mental states ... at different points in time*" (Regenwetter et al., 2011, p. 43). Here, the "*mental states*" can be translated into preferences (i.e., rank orders of the given alternatives from the most preferred to the least preferred), and decision makers are assumed to have probability distributions over such preferences so that they would randomly choose one and make choices according to the chosen preferences at the moment. In other words, decision makers would jump between multiple preferences for every task, even when the same decision task repeatedly appears. In this section, we review the mixture models of transitivity and lexicographic semiorders. Specifically, we examine how to derive inequalities for both theories, so that we can make the theories readily testable against empirical data.

Mixture model of transitivity The way the mixture model accounts for choice variability clearly contrasts with how the error model accounts for choice variability, and such difference results in different modeling approaches for transitivity. The error model assumes a core transitive preference and error, where the error part is supposed to account for all the observed choice variability. The mixture model, however, states that if the observed choices are to satisfy transitivity, those choices should come from transitive preferences without error although decision makers are allowed to switch among the transitive preferences. As long as the choices stay within the set of transitive preferences, they are said to satisfy transitivity.

Let $\mathcal{C}$ be a set of choice alternatives, and for an illustrative purpose, set the number of the alternatives to 3, $\mathcal{C} = \{A,\ B,\ C\}$. The task is a series of pairwise comparisons. The unique gamble pairs of Set $\mathcal{C}$ are A and B, B and C, and A and C. Decision makers are then asked to make choices for these three gamble pairs multiple times. Since the choices are variable, we need to model choice probabilities rather than algebraic preference relations when testing transitivity. Let $P(A, B)$ be the probability of choosing A over B when the pair A and B is present and $\mathcal{T}$ be the set of all transitive preferences on the set $\mathcal{C}$, then the mixture model computes $P(A, B)$ as follows:

$$P(A, B) = \sum_{\{\succ \text{ that indicates } A \succ B;\ \succ \in \mathcal{T}\}} P_{\succ}, \tag{2.6}$$

where $\succ$ represents preferences among the given alternatives and $P_{\succ}$ represents the probability of each preference getting chosen. In our setting, for example, $\succ$ refers to each of the preferences that can be made out of the three alternatives, namely, ABC, ACB, BAC, BCA, CAB, CBA, $ABCA$, and $ACBA$. Note that the last two preferences, $ABCA$ and $ACBA$, are intransitive, thus those intransitive preferences will have zero probability mass.

In order to understand how the mixture model works, consider computing $P(B, C)$ using Equation 2.6. There are three alternatives in Set $\mathcal{C}$, so the possible transitive preferences are ABC, ACB, BAC, BCA, CAB, and CBA. Among the six transitive preferences, ABC, BAC, and BCA are the ones that indicate B $\succ$ C, and thus, $P(B, C)$ can be obtained by summing up the probabilities of the preferences ABC, BAC, and BCA. If the decision maker puts equal probability mass across all the transitive preferences, $P(B, C)$ will be computed by $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = 0.5$. But, the mixture model doesn't require any parametric assumptions about the probability distribution over $\mathcal{T}$ (Regenwetter et al., 2010, 2011). Hence, different individuals may employ different probability distributions over the collection of transitive preferences and end up having different choice probabilities.

Mathematically, the way the mixture model computes choice probabilities can be viewed as computing a weighted average, or a convex combination, of different states of preference. Indeed, if we require preferences not just to be transitive, but also to be strict linear order, that is, we force the choices to be binary with no indifference allowed, the mixture model of transitivity can be represented by a convex polytope (Regenwetter et al., 2010). In other words, if observed choices satisfy transitivity, they should be located inside the polytope, where each choice probability is a convex combination of transitive preferences. Note that the polytope resides within the space of binary choice probabilities, so we obtain such a convex polytope by imposing constraints over choice probabilities in the form of inequalities. This effectively means that defining a mixture model of transitivity is equivalent to a problem of finding a set of inequalities that characterize the convex polytope (see Regenwetter et al., 2010, 2011, for more discussion).

Although it's not easy, if not impossible, to obtain a minimal set of inequalities for the mixutre model of transitivity when the number of alternatives is large (i.e., $|\mathcal{C}| > 5$), it's feasible to obtain such inequalities when the number of alternatives is relatively small (i.e., $|\mathcal{C}| \leq 5$). First, if the task imposes the 2-alternative forced choice paradigm on decision makers, they must choose one between a pair of gambles offered, so a binary choice probability, say $P(A, B)$, and its complement, $P(B, A)$, sum to 1, i.e., $P(B, A) + P(A, B) = 1$. Also, the mixture model (Equation 2.6) itself implies constraints on choice probabilities, known as the *triangle inequalities* (in case of testing strict linear orders, the *triangle inequalities* are necessary and sufficient conditions for the mixture model of strict linear orders). Putting them together yields the following non-redundant minimal set of inequalities that completely characterize the mixture model of transitivity when $|\mathcal{C}| \leq 5$:

$$0 \leq \Pr(A, B) \leq 1, \tag{2.7}$$

$$\Pr(A, B) + \Pr(B, A) = 1, \tag{2.8}$$

$$\Pr(A, B) + \Pr(B, C) - \Pr(A, C) \leq 1. \tag{2.9}$$

Equation (2.7) comes from the definition of probability, and Equation (2.8) is implied by the 2-alternative forced choice setting. Equation 2.9 is the triangle inequalities, implied by the mixture model as mentioned above. When the number of alternatives is 5, the triangle inequalities yield a set of 20 individual inequalities. When the number of alternatives is greater than 5, however, the number of inequalities increases to the point where testing the model is not practically possible. Given that most preferential choice experiments, including Tversky's (1969), have been done with at most 5 alternatives (e.g., Cavagnaro & Davis-Stober, 2014; Davis-Stober et al., 2015), I conclude that the above system of inequalities completes the mixture model of transitivity.

Mixture model of lexicographic semiorder The mixture model of lexicographic semiorder relies on the same assumption that underlies the mixture model of transitivity: Variable choices come from uncertainty in preferences, not from errors. That is, those whose choices conform to the mixture model of lexicographic semiorders may change their preferences from time to time during a decision-making session, but at each decision-making moment, they must make decisions according to one of the lexicographic semiorders. Unlike transitivity, however, lexicographic semiorders take into account the information about the attributes of alternatives. Thus, if the given alternatives are composed of large number of attributes, lexicographic semiorders could yield a lengthy list of semiorders, which may keep us from building a mixuture model of it. For that reason, here I consider only two-attribute alternatives, like monetary gambles, for a mixture model of lexicographic semiorders.[1]

A lexicographic semiorder is a sequential application of semiorders to the attributes, and semiorders are binary relations that exploit the concept of "just noticeable difference" (Luce, 1956). In other words, semiorders are the ones that determine how to compare two alternatives on the attribute under consideration; the lexicographic part just states the sequence of such semiorders. So, in order to understand lexicographic semiorders, we should examine semiorders first. I will follow the mathematical notations of Davis-Stober (2010, 2012) to define semiorders. Let S be a strict semiorder on the given set $\mathcal{C}$ of choice alternatives. Then, there exists a real-valued function $g(\cdot)$, e.g., utility function, defined on $\mathcal{C}$ such that, for all $x, y \in \mathcal{C}$,

$$xSy \Leftrightarrow g(x) > g(y) + q, \tag{2.10}$$

where q is a non-negative constant. Equation 2.10 shows the core idea of semiorder via constant q, which serves as a threshold for recognizing the difference between x and y. That is, only the difference greater than q will be recognized and affect the process of decision-making. The numerical function and threshold of a semiorder vary across different attributes, and thus, we need to require different semiorders for different attributes if there are more than one attribute to consider.

Lexicographic semiorders then employ semiorders to describe how one arrives at final decisions, according to one's priority among the attributes. For example, consider a choice between a pair of monetary gambles. Each of the monetary gambles is defined on two attributes: probability of winning and payoff. Let S_1 and S_2 be semiorders on probability of winning and payoff, respectively, then a binary relation of lexicographic semiorder, denoted by $\succ_{LS}$, states that: for all $x, y \in \mathcal{C}$,

[1] Lexicographic semiorders on two-attribute alternatives are termed "*simple* lexicographic semiorders" in Davis-Stober (2010).

$$x \succ_{LS} y \Rightarrow \{xS_1y\} \text{ or } \{\sim (xS_1y \text{ or } yS_1x) \text{ and } xS_2y\},$$

if the decision maker prioritizes probability of winning, and

$$x \succ_{LS} y \Rightarrow \{xS_2y\} \text{ or } \{\sim (xS_2y \text{ or } yS_2x) \text{ and } xS_1y\},$$

if the decision maker prioritizes payoff. We are then able to find all lexicographic semiorders by varying the level of threshold of each attribute within each set of semiorders. Davis-Stober (2012) defined a mixture model on those lexicographic semiorders, which, as with the mixture model of transitivity, allows decision makers to switch between preferences during a decision-making session. Let $\mathcal{LS}$ be the set of all lexicographic semiorders on $\mathcal{C}$, then $P(x, y)$ is given by:

$$P(x,y) = \sum_{\{\succ_{LS} \text{ that indicates } x \succ_{LS} y;\ \succ_{LS} \in \mathcal{LS}\}} P_{\succ_{LS}}, \tag{2.11}$$

where $P_{\succ_{LS}}$ denotes the probability of a decision maker staying in the state of a certain lexicographic semiorder.

Note that the above mixture model of lexicographic semiorders does not allow a decision maker to change the order of the attributes, in which the decision maker examines. That is, if a decision maker decides to consider probability of winning first and payoff next, he or she must stick to that order throughout the entire decision-making session, according to the model. Hence, we need to consider a different mixture model for every different sequence of the attributes in order to account for full complexity of data. Consider monetary gambles. There are two possible sequences of the attributes, in which a decision maker examines, i.e., probability of winning first, then payoff, and payoff first, then probability of winning. Therefore, we need to assume two different mixture models of lexicographic semiorders, when it comes to modeling choices for monetary gambles.

The mixture model of lexicographice semiorder in Equation 2.11 states binary choice probabilities in similar fashion to the mixture model of transitivity (Equation 2.6). In other words, each of the choice probabilities that conform to the mixture model of lexicographic semiorders can be represented as a convex combination of compatible lexicographic semiorders; that is, the mixture model (Equation 2.11) forms a convex polytope, defined as a convex hull of all lexicographic semiorders (Davis-Stober, 2012). Davis-Stober (2012) has derived a minimal set of linear inequalities, or facet-defining inequalities, of such convex polytopes for any finite n alternatives. When $n = 5$ as in Tversky's (1969) gamble experiment, for example, this system generates 39 non-redundant linear inequalities (Davis-Stober et al., 2015) that completely characterize the mixture model of lexicographic semiorders (see Davis-Stober, 2012, for a thorough discussion).

2.3 Statistical inferences

In this section, I discuss statistical inferences of choice models I've discussed thus far. This section serves to fulfill the second part of the Luce and Narens' challenge, where they emphasize the need of developing statistical models of representational theories and appropriate methods to carry out statistical tests of those models against empirical data. As we've seen above, statistical models of representational theories (e.g., stochastic transitivity) are different from other models that we'd normally encounter, such as regression models, or analysis of variance type models. Statistical models of representational theories usually operate on binary choice probabilities, by imposing a set of order constraints on this parameter space. Thus, the resulting statistical models of representational theories usually take a form of collections of inequality constraints over binary choice probabilities, and statistical inferences about those models require non-trivial amount of time and effort, if we do it from scratch.

Yet, it is still important to conduct appropriate statistical tests of representational theories against data, because representational theories provide core building blocks for numerous decision theories. For example, transitivity is a necessary condition of all utility-based economic theories, so whether the given data satisfy transitivity must be checked before any utility-based theory is applied. In what follows, I consider a number of ways to conduct statistical inferences about representational theories. The approaches are largely divided into the "frequentist" way and Bayesian way. I first start with the frequentist way of statistical inferences.

2.3.1 Frequentist statistical inference

Most frequentist statistical inferences of representational theories rely largely on maximum likelihood estimation (Regenwetter et al., 2010). Recall that in the field of decision theory, observed choices are considered realizations of inner state of true preference, so interest lies in estimating preference state that gives rise to observed choices, rather than the observed choices *per se*. From this point of view, one is able to construct a pool of preference states with given alternatives. Consider a set of three alternatives $\mathcal{C} = \{A,\ B,\ C\}$. We are interested in binary preferences, so pair-wise comparisons of $\mathcal{C}$ are considered: A and B, A and C, and B and C. For each pair-wise comparison, we write 1 if true preference indicates that the left alternative is preferred to the right; 0, otherwise. For example, between the pair A and B, if A is preferred to B, then we write the preference as 1; then, the preference state $A \succ B \succ C$ can be written as 111, where the first number represents $A \succ B$, the second number represents $A \succ C$, and the third number represents $B \succ C$. In this way, we can write a set $\mathcal{S}$ of whole preference states as combinations of $\{0, 1\}$, producing $2^{\binom{3}{2}} = 8$ preference states. Decision makers are then assumed to take one of these preference states randomly, and make choices based on the chosen state. In this case, the main interest of estimation centers on the probability P_s of preference state

s. Let $\boldsymbol{P} = (P_s)_{s=1,\ldots,|S|}$ and $\boldsymbol{n} = (n_s)_{s=1,\ldots,|S|}$, where n_s is the number of preference state s having been occurred out of sample size N. Then, the likelihood function of $\boldsymbol{P}$ is as follows:

$$\mathcal{L}(\boldsymbol{P}|N,\ \boldsymbol{n}) = \prod_{s\in S} \binom{N}{n_s} P_s^{n_s},$$

$$\text{where } \sum_{s\in S} P_s = 1, \text{ and } \sum_{s\in S} n_s = N.$$

One downside of the above approach is that the number of free parameters (i.e., P_s) increases exponentially with the number of alternatives, quickly reaching to the point where its estimation is no longer practical. For example, when the number of alternatives is 5, the number of preference states increases to $2^{\binom{5}{2}} = 1024$, resulting in $1024 - 1 = 1023$ free parameters, i.e., P_s. Thus, it is typical not to deal with the preference states as is, but instead to use binary choice probabilities of the given alternatives. Consider again the previous example with three alternatives $\{A,\ B,\ C\}$. We are now interested in estimating binary choice probabilities of the given alternatives, $P_{AB},\ P_{AC},$ and P_{BC}, where P_{AB} is the probability of choosing A when the pair A and B is present. Let $\boldsymbol{P}^{\mathcal{B}} = (P_{AB},\ P_{AC},\ P_{BC})$ and $\boldsymbol{n}^{\mathcal{B}} = (n_{AB},\ n_{AC},\ n_{BC})$, where n_{AB} is the number of observed choices made for A when the pair A and B is present. Then, the likelihood function of $\boldsymbol{P}^{\mathcal{B}}$ is as follows:

$$\mathcal{L}(\boldsymbol{P}^{\mathcal{B}}|N,\ \boldsymbol{n}^{\mathcal{B}}) = \prod_{k\in\{AB,AC,BC\}} \binom{N}{n_k} P_k^{n_k}(1-P_k)^{N-n_k}.$$

Now the number of free parameters is reduced from 7 to 3 when the number of alternatives is 3. The advantage of this approach is even more dramatic when the number of alternatives is 5; the number of preference states that five alternatives make is 1024, which produces 1023 free parameters, but the binary choice probability approach produces only $\binom{5}{2} = 10$ free parameters to be estimated. This is exactly what Tversky (1969) employed to test transitivity of his data. He first selected specific preference state, denoted by $\mathcal{M}_1$, to put to the test against empirical data, which can be achieved by imposing a set of constraints over the binary choice probabilities. Then, the unrestricted model $\mathcal{M}_0$ is defined by imposing no constraints over the entire parameter space. As $\mathcal{M}_1$ is nested within $\mathcal{M}_0$, Tversky performed a likelihood-ratio test of these two models, which gives a test statistics $\Lambda(\mathcal{M}_1,\ \mathcal{M}_0)$ that asymptotically follows the chi-square distribution with the degrees of freedom equal to the number of constrained parameters:

$$\Lambda(\mathcal{M}_1,\ \mathcal{M}_0) = -2ln\left[\frac{\sup_{\boldsymbol{P}^{\mathcal{B}} \in \mathcal{M}_1} \mathcal{L}(\boldsymbol{P}^{\mathcal{B}})}{\sup_{\boldsymbol{P}^{\mathcal{B}} \in \mathcal{M}_0} \mathcal{L}(\boldsymbol{P}^{\mathcal{B}})}\right],$$

where the sup notation refers to the supremum of the argument.

Since $\mathcal{M}_1$ is nested within $\mathcal{M}_0$, the maximum likelihood of $\mathcal{M}_1$ cannot exceed the maximum likelihood of $\mathcal{M}_0$. So the likelihood ratio test statistics $\Lambda(\mathcal{M}_1,\ \mathcal{M}_0)$ is bounded between 0 and 1. The rationale behind this test statistics is that if the data were actually arisen from $\mathcal{M}_1$, its maximum likelihood should not be much different from that of the unrestricted model $\mathcal{M}_0$, resulting in the test statistics close to 1. Only when the maximum likelihood of $\mathcal{M}_1$ is significantly smaller than that of $\mathcal{M}_0$, we can reject the null hypothesis, providing evidence that the data are not likely arisen from $\mathcal{M}_1$. In this way, Tversky (1969) tested the weak stochastic transitivity and lexicographic semiorder theories against empirical data, and found dominant evidence for intransitivity of preferences among the participants.

However, one major criticism about his approach for testing transitivity is that Tversky (1969) considered only a few preference states for testing transitivity. The number of alternatives used in his experiment was five, so the number of the preference states that satisfy transitivity is $5! = 120$. Of which, he only considered the most obvious transitivity patterns, $A \succ B \succ C \succ D \succ E$ and $E \succ D \succ C \succ B \succ A$.

Also, he considered only one particular way of making lexicographic semiorders, where participants would choose the one with higher payoff between a pair of adjacent gambles, but choose the one with higher probability of winning between a pair of extreme gambles. Iverson and Falmagne (1985) pointed out that if the analysis allowed participants to have different transitivity patterns, the data would have failed to reject the null hypothesis of transitivity. To address the problem of Tversky's analysis, Birnbaum and Gutierrez (2007) suggested a new statistical technique, where multiple preference states are considered for the analysis. But they used only three gambles A, C, and E out of five to form preference states, leaving gambles B and D out. I suspect that since including all five gambles results in 1024 preference states, they may have had to limit their analysis only to gambles A, C, and E, so they could keep the number of states manageable. Although Birnbaum and Gutierrez included more than a few preference states in the analysis, I doubt that we could call their approach a general method for testing representational theories.

Approaching for preference states without a system looks intimidating, as one needs to deal with so many preference states. But, as we've discussed in the previous section, statistical models of representational theories can be represented via groups of inequalities over binary choice probabilities. In this case, the parameter space that satisfies those inequalities can be recast as geometrical objects, known as convex polytopes (Regenwetter et al., 2010), and observed choices can directly be compared against the resulting convex polytopes. The problem is, however, when the observed choices are not inside the polytopes, the point estimates of the parameters of interest will lie on the boundary of the polytopes, which no longer ensures the likelihood ratio test statistics $\Sigma(\mathcal{M}_1, \ \mathcal{M}_0)$ to follow the asymptotic chi-square distribution (Davis-Stober, 2009; Iverson & Falmagne, 1985; Regenwetter et al., 2011). Acknowledging the problem, Davis-Stober (2009) generalized the Iverson and Falmagne's (1985) approach and developed a statistical test of inequality-induced models, such as weak stochastic transitivity or lexicographic semiorder error model. This approach employs the asymptotic chi-bar-square $\bar{\chi}^2$ distribution, with its weights approximated by the geometric structure of the neighborhood of the maximum likelihood estimates.

The approach developed by Davis-Stober (2009) provides a general statistical testing framework for axiomatic representational theories. It allows us to evaluate whether the observed choices located outside the mathematical model are due merely to the sampling errors, or due to significant violations of the given model (Regenwetter et al., 2011). This approach also allows us to test general types of order-constrained hypotheses, including both the error and mixture models of transitivity and lexicographic semiorders. Because of its generality, the new approach has been implemented in the original version of QTest (2014), a public domain software package, as its main statistical testing algorithm. Despite its effectiveness, however, Davis-Stober's approach has a number of limitations: first, its test statistics requires large sample sizes to exploit asymptotic distributions, second, it doesn't quantify evidence for the given model, only producing the goodness-of-fit measure, and third, it doesn't provide model selecting tools. Such limitations inspire the development of QTest 2.1 (Zwilling et al., 2019), which relies heavily on Bayesian statistics. In the following section, I review the Bayesian statistical inferences.

2.3.2 Bayesian inferences

Bayesian basics

Bayesian statistics got its name from the Bayes' theorem. Bayesian statistics rely heavily on the Bayes' theorem to make statistical inferences about any quantity of interest to researchers, such as parameters, hypotheses, etc. The Bayes' theorem states that:

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})},$$

where $\boldsymbol{\theta}$ is the parameters of substantive interest (or could be the hypothesis of interest), and D is observed data. Then, $p(D|\boldsymbol{\theta})$ represents the likelihood function, $p(\boldsymbol{\theta})$ represents prior density of $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta}|D)$ represents posterior density of $\boldsymbol{\theta}$. Note that the denominator of the Bayes' theorem is a constant with respect to $\boldsymbol{\theta}$, ensuring that $p(\boldsymbol{\theta}|D)$ integrates to 1. Thus, the Bayes' theorem is frequently expressed as follows:

$$p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{2.12}$$

where the constant of proportionality is

$$\left[\int_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})\right]^{-1}.$$

Equation (2.12) is so important and fundamental for Bayesian inferences that its close examination reveals the essence of the Bayesian approach. First, $p(D|\boldsymbol{\theta})$ represents the likelihood function that carries the same meaning as in the frequentist inferences, although its notation has changed. The prior density $p(\boldsymbol{\theta})$ formulates one's prior knowledge or beliefs with respect to the parameters $\boldsymbol{\theta}$ before the data are observed. And the posterior density $p(\boldsymbol{\theta}|D)$ formulates one's updated knowledge or beliefs about $\boldsymbol{\theta}$ after seeing the data. Thus, the chief interest of Bayesian statistics lies in updating one's prior knowledge about the parameters of substantive interest in light of data, producing the posterior beliefs about the parameters. This core idea is depicted in the following schemata of Bayesian inferences:

$$\begin{array}{ccccc} \text{Prior beliefs about } \boldsymbol{\theta} & \rightarrow & \text{Observed data} & \rightarrow & \text{Updated beliefs about } \boldsymbol{\theta} \\ p(\boldsymbol{\theta}) & & p(D|\boldsymbol{\theta}) & & p(\boldsymbol{\theta}|D) \end{array}$$

One key difference between the Bayesian approach and the classical (or frequentist) approach is the assumption about parameters. From the perspective of frequentist approach, the parameters $\boldsymbol{\theta}$ are fixed at the "true" values, which are assumed to generate the data we observe. Researchers then, seek to estimate the true values of the parameters from the data at hand, yielding an estimate of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$. The estimate $\hat{\boldsymbol{\theta}}$ varies with the data at hand, so one can obtain a sampling distribution of $\hat{\boldsymbol{\theta}}$, by assuming repeated sampling of the target population. It's this sampling distribution that most frequentist hypothesis testing tools rely on, including p-values, and confidence intervals.

The Bayesian approach, on the other hand, regards the parameters $\boldsymbol{\theta}$ as random variables, not fixed quantities. This is because the Bayesian approach adopts subjective probability and the parameters $\boldsymbol{\theta}$ being random reflect that researchers are not certain about its quantity. In this sense, it makes sense to state that one's prior beliefs can be updated in light of data, with the posterior density $p(\boldsymbol{\theta}|D)$ representing the updated beliefs about $\boldsymbol{\theta}$ after seeing the data.

One of benefits of the Bayesian approach is that the posterior density of Bayesian statistics provides intuitive interpretations of statistical inferences. When it comes to testing hypotheses, the frequentist statistics rely heavily on p-values. If a p-value is less than an arbitrary small value, usually set to 0.05, one would claim to reject the null hypothesis, which in turn provides evidence for the alternative hypothesis. Aside from many potential problems with relying solely on p-values for statistical inferences (see Wagenmakers, 2007, for a thorough discussion on this topic), it is the interpretation that usually makes researchers confused about their findings. Suppose that a researcher is interested in testing the null hypothesis of one of regression coefficients equal to 0, say $H_0 : \beta_h = 0$. If a frequentist statistical test of the hypothesis yields a p-value less than 0.05, the researcher can reject the null hypothesis at $\alpha = 0.05$ (Type-I error rate). Since the p-value is computed based on the sampling distribution of $\hat{\beta}_h$, the interpretation of the result goes: given that the null hypothesis is true, the probability of getting the same result as the one we obtain or more extreme results if the same experiment is conducted over and over again on the samples from the same population is less than 0.05. On the other hand, Bayesian inferences rely on the posterior density of β_h conditional on observed data, of which the interpretation goes: Given the data we observe, what's the probability of β_h being equal to 0? This way of interpretation is strikingly similar to what we want to say about the hypothesis testing, and Bayesian methods allow us to interpret the results the way we want.

Despite its intuitive approach for statistical inferences, however, the Bayesian statistics had not been widely used due to the difficulty in evaluating posterior density. In practice, it is easy for the likelihood function to get complicated with multiple parameters. As one adds more parameters to the model, the joint posterior density quickly becomes intractable, making its use in practice impossible. This is not just limited to complicated models. For example, regression models for binary choice data, i.e., logit or probit models, are known to have no closed form, hence no analytical solutions exist for the maximum likelihood estimates (McCullagh & Nelder, 1989). Deriving the posterior density of this model is not any better; as Jackman (Jackman, 2009) demonstrated, no conjugate priors exist for the likelihood function of these models, yet the denominator of the posterior density involves the multi-dimensional integration, making the posterior density largely intractable. Bayesian methods had been out of reach from most researchers due mainly to this computational difficulty. But, with the recent development in statistical techniques, coupled with the massive increase in computer power, now the Bayesian approach is made feasible for almost all statistical problems. The way to deal with such intractable posterior density is via *sampling*, not via *optimizing* (Jackman, 2000). That is, Bayesian methods tackle inferential problems by simulating posterior density using newly developed statistical techniques, called the Markov chain Monte Carlo (MCMC) algorithm (Gelfand & Smith, 1990).

The MCMC sampling algorithm combines two different ideas: Markov chain and Monte Carlo. Each of these two ideas is revolutionary in statistics on its own, but when they are combined together, it entirely transforms the way we approach the Bayesian inferences. Equipped with MCMC algorithms, the Bayesian methods are now able to tackle virtually any statistical problems and generate statistical inferences even for the problems that have been considered intractable. The core idea is that if we can sample from the target distribution (e.g., a posterior density in case of Bayesian inferences), any features about the distribution can be learned by the samples (Jackman, 2009). This is the core principle of Monte Carlo methods, which is well summarized by the following quote:

> anything we want to know about a random variable θ can be learned by sampling many times from $f(\theta)$, the density of θ (Jackman, 2009, p. 133).

Even if it is not possible to sample directly from the target distribution, we can still obtain samples of the target distribution, thanks to the Markov chain theory. That is, if we construct a proper Markov chain on the parameter space, the chain will travel within the parameter space such that frequencies of the visited locations are proportional to the probability of the locations under a density of interest, i.e., a posterior density. Such Markov chains are said to be *ergodic*. In this case, we can store the trajectory of the chain and treat those values as samples from the posterior density of interest. Then, Monte Carlo methods will help us make statistical inferences about the posterior density using the sequence of values generated by the chain. As of now, there are a number of numerical algorithms that generate ergodic Markov chains, such as Metropolis-Hastings algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), and the Gibbs sampler (Casella & George, 1992; Gelfand, Smith, & Lee, 1992). Those algorithms ensure that the constructed chain is ergodic, so that it produces a series of parameter draws that eventually converge to samples from the exact posterior density we aim to obtain.

In what follows, I review Bayesian statistical testing tools, the Bayesian p-values, Deviance Information Criterion (DIC) and Bayes factors, which exploit these MCMC techniques.

Bayesian p-value

In this section, I discuss the Bayesian p-value, which can be obtained by exploiting posterior predicted samples (Meng, 1994; Myung et al., 2005). The Bayesian p-value can be employed to test axiomatic representational theories, such as transitivity, against data, hence this statistic is particularly related to the current study in my opinion.

The Bayesian p-value is the Bayesian equivalent of the frequentist p-value, so it provides a measure of how well the model of substantive interest fits against empirical data. To compute the Bayesian p-value, one needs to compute the posterior predictive density, which describes the density of the "future" data, predicted by the model of interest if the same experiment that generates observed data is undertaken repeatedly. Then, the Bayesian p-value is computed via comparing the observed data with the hypothetical data predicted by the model, which provides an effective tool for evaluating the model fit to the data.

First, we start with the posterior density of $\boldsymbol{\theta}$, because the posterior predictive density can be obtained via posterior samples of $\boldsymbol{\theta}$ using the Monte Carlo method. Let $\Omega_{\mathcal{M}}$ is the parameter space constrained by the model $\mathcal{M}$. Then, the posterior density $\boldsymbol{\theta}$ conditional on the data $\boldsymbol{y}$, $p(\boldsymbol{\theta}|\boldsymbol{y})$, is given by:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Omega_{\mathcal{M}}} p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

One thing worth mentioning here is the way the authors set the priors for $\boldsymbol{\theta}$, denoted by $p(\boldsymbol{\theta})$. When it comes to modeling representational theories, the parameter vector $\boldsymbol{\theta}$ consists of binary choice probabilities for all possible pairs of the given alternatives. Consider, for example, a set $\mathcal{C}$ of three alternatives, $\mathcal{C} = \{A,\ B,\ C\}$. Then, $\boldsymbol{\theta} = (\theta_{AB}, \theta_{AC}, \theta_{BC})$, where θ_{AB} is the probability of choosing A over B. Since probability is bounded between 0 and 1, the parameter space of $\boldsymbol{\theta}$ is a n-dimensional unit hypercube (when $|\mathcal{C}| = 3$, the parameter space is a unit cube), where every point represents a unique preference structure. As Iverson and Falmagne (1985) observed, a statistical model of representational theories can be expressed as a set of constraints over this parameter space. Let Θ be the unconstrained parameter space of $\boldsymbol{\theta}$, i.e., $\Theta = [0, 1]^n$, where its dimensionality n is given by the number of unique pairs that the given alternatives make, and $\Omega_{\mathcal{M}}$ the parameter space constrained by the model $\mathcal{M}$. Then, by definition, $\Omega_{\mathcal{M}} \subseteq \Theta$, and statistical inferences are made in terms of $\Omega_{\mathcal{M}}$. Myung et al. (2005) made use of $\Omega_{\mathcal{M}}$ to infer the posterior density of $\boldsymbol{\theta}$ by limiting its prior to $\Omega_{\mathcal{M}}$. Specifically, let $h(\theta)$ be a probability measure on $\Omega_{\mathcal{M}}$, then the prior $p(\boldsymbol{\theta})$ is given by:

$$p(\boldsymbol{\theta}) = \frac{h(\boldsymbol{\theta})}{\int_{\Omega_{\mathcal{M}}} h(\boldsymbol{\theta}) d\boldsymbol{\theta}},$$

where $h(\boldsymbol{\theta})$ allows no probability mass outside the region $\Omega_{\mathcal{M}}$.

With the prior specified above, one can build a proper Markov chain using one of the MCMC techniques mentioned above. Particularly, the authors used the Gibbs sampler to obtain samples $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, ..., \theta_n^{(t)}),\ t \in \{1, ..., T\}$ from the posterior density $p(\boldsymbol{\theta}|\boldsymbol{y})$ (see Myung et al., 2005, for a detailed account of the sampling steps). Then, they employed the Monte Carlo techniques to estimate the posterior predictive density using the posterior samples. Let $\boldsymbol{y}^{\text{pred}}$ be the future data predicted by model $\mathcal{M}$. The posterior predictive density $p(\boldsymbol{y}^{\text{pred}}|\boldsymbol{y})$ is then given by:

$$p(\boldsymbol{y}^{\text{pred}}|\boldsymbol{y}) = \int_{\Omega_{\mathcal{M}}} p(\boldsymbol{y}^{\text{pred}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}.$$

The Monte Carlo methods allow for generating samples from $p(\boldsymbol{y}^{\text{pred}}|\boldsymbol{y})$ using posterior samples of $\boldsymbol{\theta}$. Specifically, at each iteration t, a sample of the "future" data $D_1^{(t)}$ is drawn from a binomial distribution with its success parameter set to $\theta_1^{(t)}$. Note that the data for each pair are generated independently, i.e., $D_i^{(t)} \sim \text{Binomial}(\theta_i^{(t)}),\ t \in \{1, ..., T\}$, with each sample serving as future data from the correct posterior predictive density $p(\boldsymbol{y}^{\text{pred}}|\boldsymbol{y})$. This way of estimating a posterior predictive distribution, instead of using just a point estimate of $\boldsymbol{\theta}$, such as $E(\boldsymbol{\theta}|\boldsymbol{y})$, is recommended, because the resulting posterior predictive samples take into account the uncertainty with respect to $\boldsymbol{\theta}$, as well as to the sampling process (Gelman et al., 2013).

In order to obtain the Bayesian p-value, the authors again employed the Monte Carlo method that utilizes the generated posterior predictive samples. Since the Bayesian p-value is a measure of discrepancy between the data at hand and the future data predicted by the model $\mathcal{M}$ of interest, the generalized Pearson chi-square discrepancy function is used, which is defined by:

$$\chi^2(\boldsymbol{y};\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{(y_i - N_i\theta_i)^2}{N_i\theta_i},$$

where N_i is sample size of the i-th pair, $i \in \{1, ..., n\}$. Then, the corresponding p-value is given by:

$$\text{Bayesian } p\text{-value} \equiv \Pr\{\chi^2(\boldsymbol{y}^{\text{pred}};\boldsymbol{\theta}) \geq \chi^2(\boldsymbol{y};\boldsymbol{\theta})\}.$$

Finally, the above Bayesian p-value can be estimated by applying the Monte Carlo method as follows:

$$\frac{1}{T}\sum_{t=1}^{T} I(\chi^2(\boldsymbol{y}^{\text{pred}(t)};\boldsymbol{\theta}^{(t)}) \geq \chi^2(\boldsymbol{y};\boldsymbol{\theta}^{(t)})),$$

where $I(\cdot)$ is a function that returns 1 if its argument is true, 0 otherwise. And $\boldsymbol{y}^{\text{pred}(t)}$ represents posterior predictive samples from $p(\boldsymbol{y}^{\text{pred}}|\boldsymbol{y})$ and $\boldsymbol{\theta}^{(t)}$ represents posterior samples from $p(\boldsymbol{\theta}|\boldsymbol{y})$.

As defined above, the Bayesian p-value is the probability of the χ^2 value of the future data predicted by $\mathcal{M}$ being greater than or equal to the χ^2 value of the data at hand. So, a large Bayesian p-value indicates that the model $\mathcal{M}$ provides adequate fit to the observed data; a small Bayesian p-value, on the other hand, indicates a lack of fit of $\mathcal{M}$ against the observed data. The authors illustrated a series of examples of applications of the Bayesian p-value, including test of the weak stochastic transitivity (WST), moderate stochastic transitivity (MST), and strong stochastic transitivity (SST) against Tversky's (1969) data. The Bayesian p-value has also been successfully applied to test of the axioms of conjoint measurement (Karabatsos & Batchelder, 2003) and the axioms of cultural consensus models (Karabatsos & Sheu, 2004).

Deviance information criterion (DIC)

As Myung et al. (2005) pointed out, one needs to be cautious about the interpretation of the Bayesian p-value, as it doesn't represent the probability that the given model is true. It only provides the information as to how well the given model fits against the observed data. A small p-value can provide evidence of lack of fit of the model against the data, but a large p-value cannot be used as evidence for selecting among candidate models. Myung et al. recommended to use the Bayesian p-value as a sort of screening tool, so that we could screen out ill-fitting models from further analyses. Then, we can apply a model selection method to the revised list of models to produce the best-fitting model. The authors suggest the Deviance Information Criterion (DIC) for the model selection analysis.

The DIC is a measure of generalizability of the given model, which can be estimated via a tradeoff between goodness-of-fit and complexity of the model. Here, the generalizability of a model, as the authors put it, represents prediction accuracy, thus the model selection problem can be stated as follows:

> Given a set of two or more axioms $\mathcal{M}_k,\ k \in \{1, ..., M\}$ identify that axiom with the highest generalizability (prediction accuracy) over future observations of the same experiment that generated the current data set $\boldsymbol{y}$ (Myung et al., 2005, p. 210; the mathematical notations have been edited to match the context).

The DIC consists of two components: the lack-of-fit measure, and the penalty term with respect to model complexity. The computation of lack-of-fit of the DIC relies on the deviance discrepancy function $D(\boldsymbol{\theta})$ (McCullagh & Nelder, 1989), which is given by:

$$D(\boldsymbol{\theta}) = 2\sum_{i=1}^{n}\left[y_i \text{log}\left(\frac{y_i + 1/2}{N_i\theta_i + 1/2}\right) + (N_i - y_i)\text{log}\left(\frac{N_i - y_i + 1/2}{N_i - N_i\theta_i + 1/2}\right)\right],$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ is a posterior sample from $p(\boldsymbol{\theta}|\boldsymbol{y})$.

The model complexity is then computed via the effective number of parameters p_D with respect to the given model (Spiegelhalter, Best, Carlin, & van der Linde, 2002), which is given by:

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}),$$

where $\overline{D(\boldsymbol{\theta})} = \frac{1}{T}\sum_{t=1}^{T} D(\boldsymbol{\theta}^{(t)})$ and $D(\bar{\boldsymbol{\theta}}) = D(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\theta}^{(t)})$. Hence, p_D quantifies the difference between the average of the deviance functions and the deviance function evaluated at the average of the posterior samples. As Spiegelhalter and his colleagues (2002) elaborated, such a difference provides an important quantity in estimating the degrees of freedom of a test. Then, the DIC is computed as follows:

$$\text{DIC} = D(\bar{\boldsymbol{\theta}}) + 2p_D.$$

The computed value of DIC represents an estimated averaged distance between the model of substantive interest and the true data-generating model. Since it measures distances, the resulting value doesn't have a special meaning to it when interpreted alone. Only when there are two or more models to be compared, the DIC provides ordinal measures for the candidate models, where smaller values of DIC indicates better generalizability of a model. Hence, the model with the smallest DIC value is preferred.

The major benefit of DIC is that its computation is easily done using the posterior samples obtained via the MCMC techniques. For Bayesian inferences, one typically generates posterior samples of parameters of interest from its posterior density using the MCMC sampling algorithm. Once the posterior samples are generated, the DIC can be almost automatically given as a byproduct of these samples. In addition, the penalty term p_D of DIC can account for the complexity of order-constrained models, which contrasts with the penalty term of the Akaike Information Criterion (AIC) that only takes into account the number of parameters in the model. Also, unlike the likelihood ratio test, DIC doesn't require candidate models to be nested within each other, providing a much flexible model selection tool. Therefore, the DIC, coupled with the Bayesian p-value, offers a complete package of Bayesian inferences.

In the following section, we discuss Bayes factor as another model selection tool. Bayes factor is different from DIC in that it doesn't rely on any externally defined function, like the deviance discrepancy function. Instead, the Bayes factor is based on the terms naturally derived from Bayes theorem, providing an intuitive and straightforward measure of evidence for competing models. The computation of Bayes factor had been considered difficult for most statistical models, but with recent developments in numerical algorithm, we are now able to obtain Bayes factors quite easily at least for axiomatic representational theories.

Bayes factor

Suppose that we have a set of data $\boldsymbol{y}$ and that we are interested in choosing a better model for data $\boldsymbol{y}$ between two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$. This is a typical model selection problem (Myung & Pitt, 1997). The Bayes factor, as a Bayesian model selection method, provides a natural solution for this problem. For simplicity, suppose that we consider only two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, and one of the two models is assumed to have arisen the data $\boldsymbol{y}$. Then, the posterior density of $\mathcal{M}_k,\ k = 1, 2$ is given by:

$$p(\mathcal{M}_k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathcal{M}_k)p(\mathcal{M}_k)}{p(\boldsymbol{y}|\mathcal{M}_1)p(\mathcal{M}_1) + p(\boldsymbol{y}|\mathcal{M}_2)p(\mathcal{M}_2)},$$

which yields the posterior odds ratio $p(\mathcal{M}_1|\boldsymbol{y})/p(\mathcal{M}_2|\boldsymbol{y})$ as follows:

$$\frac{p(\mathcal{M}_1|\boldsymbol{y})}{p(\mathcal{M}_2|\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\mathcal{M}_1)}{p(\boldsymbol{y}|\mathcal{M}_2)}\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}.$$

Since the posterior density of $\mathcal{M}_1$ represents the likelihood of $\mathcal{M}_1$ given the data $\boldsymbol{y}$, the posterior odds ratio represents the relative likelihood of $\mathcal{M}_1$ with respect to $\mathcal{M}_2$, given the data. In the same way, $p(\mathcal{M}_1)/p(\mathcal{M}_2)$ is the prior odds ratio, which represents the relative likelihood of $\mathcal{M}_1$ with respect to $\mathcal{M}_2$, before the data are given. We update our prior odds ratio into the posterior odds ratio in light of data, and so how much evidence the data provide for each model is a decisive factor for this update. And we call this evidence ratio $p(\boldsymbol{y}|\mathcal{M}_1)/p(\boldsymbol{y}|\mathcal{M}_2)$ the Bayes factor. We can rewrite the above formula in words:

$$\text{posterior odds ratio} = \text{Bayes factor} \times \text{prior odds ratio.}$$

Thus, the Bayes factor can also be viewed as the ratio of posterior odds to the prior odds. If we assume that the two competing models are equally probable *a priori*, the

Bayes factor becomes the posterior odds ratio.

One of the strengths of the Bayes factor is that its computation automatically takes into account the measures of goodness of fit and model complexity. The Bayes factor takes the evidence for each model for its computation and such evidence is computed via marginal likelihoods. Let $\boldsymbol{\theta}_{\mathcal{M}_k}$ be the parameter vector under model $\mathcal{M}_k,\ k = 1, 2$. Then, the marginal likelihood of $\mathcal{M}_k$ is defined by:

$$p(\boldsymbol{y}|\mathcal{M}_k) = \int_{\boldsymbol{\theta}_{\mathcal{M}_k}} p(\boldsymbol{y}|\boldsymbol{\theta}_{\mathcal{M}_k}, \mathcal{M}_k)p(\boldsymbol{\theta}_{\mathcal{M}_k}|\mathcal{M}_k)d\boldsymbol{\theta}_{\mathcal{M}_k}, \quad k = 1, 2.$$

The above formulation indicates that the marginal likelihood of $\mathcal{M}_k$ is obtained by integrating out its model parameters $\boldsymbol{\theta}_{\mathcal{M}_k}$. As Kass and Raftery (1995) pointed out, the Bayesian approach for model selection is different from that of the frequentist approach in that while the frequentist approach *maximizes* the likelihood, the Bayesian approach *integrates* the likelihood over the parameter space. And it's this integration part that takes into account the model complexity. Since the integral of the likelihood is evaluated over the entire parameter space, complex models, determined by the number of parameters, functional form, and range of parameters, are in general hard to score a good result in this integration (Myung & Pitt, 1997). That is, one can make a model complex enough to make the peak of its likelihood higher than other models for the given data, but aside from that region, other values in the parameter space very likely produce low likelihood as the model gets complicated, resulting in low marginal likelihoods. In this sense, Myung and Pitt (1997) expressed the marginal likelihood in words as below:

$$\text{marginal likelihood} = \frac{\text{goodness of fit}}{\text{model complexity}}.$$

The Bayes factor is the ratio of such marginal likelihoods of the two competing models:

$$\mathrm{BF}_{12} = \frac{p(\boldsymbol{y}|\mathcal{M}_1)}{p(\boldsymbol{y}|\mathcal{M}_2)} = \frac{\int_{\boldsymbol{\theta}_{\mathcal{M}_1}} p(\boldsymbol{y}|\boldsymbol{\theta}_{\mathcal{M}_1}, \mathcal{M}_1) p(\boldsymbol{\theta}_{\mathcal{M}_1}|\mathcal{M}_1) d\boldsymbol{\theta}_{\mathcal{M}_1}}{\int_{\boldsymbol{\theta}_{\mathcal{M}_2}} p(\boldsymbol{y}|\boldsymbol{\theta}_{\mathcal{M}_2}, \mathcal{M}_2) p(\boldsymbol{\theta}_{\mathcal{M}_2}|\mathcal{M}_2) d\boldsymbol{\theta}_{\mathcal{M}_2}}.$$

Computation of the Bayes factor has been made easier thanks to recently developed numerical algorithms. In case of the problem of representational theories, particularly, the Bayes factor can be easily computed due to the work by Klugkist and Hoijtink (2007). Their method is to define the model of interest as a subset of the unconstrained, encompassing model. Then, by exploiting the nesting feature of the models, the authors show that the Bayes factor can nicely be estimated by the ratio of two proportions. Let $\mathcal{M}_0$ is the unconstrained, encompassing model and $\mathcal{M}_1$ the model of interest that put a set of constraints within $\mathcal{M}_0$. The Bayes factor of $\mathcal{M}_1$ with respect to $\mathcal{M}_0$, BF_{10}, is then given by:

$$\mathrm{BF}_{10} = \frac{\int_{\boldsymbol{\theta}_{\mathcal{M}_1}} p(\boldsymbol{y}|\boldsymbol{\theta}_{\mathcal{M}_1}, \mathcal{M}_1) p(\boldsymbol{\theta}_{\mathcal{M}_1}|\mathcal{M}_1) d\boldsymbol{\theta}_{\mathcal{M}_1}}{\int_{\boldsymbol{\theta}_{\mathcal{M}_0}} p(\boldsymbol{y}|\boldsymbol{\theta}_{\mathcal{M}_0}, \mathcal{M}_0) p(\boldsymbol{\theta}_{\mathcal{M}_0}|\mathcal{M}_0) d\boldsymbol{\theta}_{\mathcal{M}_0}} = \frac{c_1}{d_1},$$

where $1/c_1$ is the proportion of the encompassing prior that is in agreement with the constraints of model $\mathcal{M}_1$ and $1/d_1$ is the proportion of the encompassing posterior that is in agreement with the constraints of model $\mathcal{M}_1$. I strongly recommend interested readers to consult the paper (Klugkist & Hoijtink, 2007) for a full derivation of the method for computing the Bayes factor.

One nice thing about this encompassing model method is that it applies for most statistical models of representational theories, because those models sit naturally inside the unit hypercube as they impose constraints over the binary choice probabilities. Even if there are multiple models to be considered, the encompassing model approach still applies as long as those models are nested within the same encompassing model. Consider, for example, a set of candidate models $\{\mathcal{M}_1, ..., \mathcal{M}_M\}$, with the same encompassing model $\mathcal{M}_0$ shared by all models. Then, we can compute Bayes factors for all candidate models against the encompassing model:

$$\mathrm{BF}_{m0} = \frac{c_m}{d_m}, \quad m \in \{1, ..., M\}.$$

Since the Bayes factor is simply a ratio of evidence of two models, it is straightforward to obtain a Bayes factor of any two of the candidate models, say $\mathcal{M}_1$ and $\mathcal{M}_2$:

$$\mathrm{BF}_{12} = \frac{\mathrm{BF}_{10}}{\mathrm{BF}_{20}}.$$

The encompassing model approach for Bayes factors has been extensively used to test representational theories, specifically for transitivity and lexicographic semiorder models (Cavagnaro & Davis-Stober, 2014; Davis-Stober et al., 2015, 2019; Davis-Stober, Park, Brown, & Regenwetter, 2016; Park et al., 2019). The encompassing method has also been implemented in the QTest 2.1 (Zwilling et al., 2019) as the main numerical algorithm for computing the Bayes factor of order-constrained models. Because of its straightforward interpretation and intuitive way to implement the measures of goodness of fit and model complexity, I employ the Bayes factor as my main model selection tool for the analyses of choice data.

Chapter 3

Diffusion model

In our every day life, we frequently observe people make choices differently. Even for the same problems, it is not hard to see different people give different answers, often making us wonder why. Such observations raise questions as to why people behave differently, or what makes them behave in that specific way. And these are the kind of questions cognitive psychologists have long been seeking to answer, and for that they study response time (RT) (Luce, 1986).

Studying RT to infer cognitive processes has a long history in cognitive psychology. One of the first attempts to examine RT for studying cognitive processes dates back to as early as 1800's, when Donders (1868, as cited in Luce, 1986) suggested the use of RT to differentiate between mental processes. His idea was, if we run pairs of well-designed experiments, where one of the experimental conditions involves an extra feature that requires a particular stage of the cognitive process of interest, then we could compute the time taken for that particular stage by comparing the RTs between different experimental conditions. Similarly, Jastrow (1890, as cited in Luce, 1986) (Jastrow, 1890) argued that if our mind is highly structured, and RTs reflect different time courses of different paths through that structure, then we could infer back the structure of our mind by examining the patterns of RTs under various experimental conditions. No matter the validity of those early assertions on RT, RT has remained as a valuable tool for the study of cognitive processes since then. Indeed, RT measurements have become a ubiquitous dependent variable in almost all fields of cognitive psychology ever since, being used for inferring various types of cognitive processes in various fields (Anders, Alario, & Van Maanen, 2016; Colonius & Marley, 2015; Gill & Prowse, 2017; Luce, 1986; Spiliopoulos & Ortmann, 2018).

Studying RT, however, is not as straightforward as it may sound. One may think he or she could just compare RTs for target tasks between experimental groups of interest, but naive comparisons of RTs might end up with misleading conclusions due to the following reasons. First, the distribution of RT is positively skewed, and so any single summary statistics of the distribution, such as mean or median, provides a poor measure of central tendency. A positively skewed distribution has a long right tail, with most of data away from its mean to the left. In such cases, the mean tends to be overestimated because of the thick right tail (Rousselet & Wilcox, 2020) and thus, is not recommended for statistic tests that rely on asymptotic properties of normal distribution, such as ANOVA. The median might be a better choice for summarizing the distribution, but when a distribution is skewed, the sample median will also be biased to some degree (Miller, 1988), not making a good choice for a test of group differences either.

Second, RT is known to have multiple sources of variability, which may not be fully considered if one blindly compares between RTs. For example, any kind of cognitive tasks (e.g., lexical decision task, numerosity discrimination task) requires one to encode the relevant information from stimuli, to process that information, and to excute a decision. Each part of the decision-making process takes time, and if those parts are not properly addressed, researchers may not be able to give a sensible account of the cognitive process of interest (Ratcliff, Thapar, & McKoon, 2004; Vandekerckhove, Verheyen, & Tuerlinckx, 2010).

Finally, the pattern of RT measuements has been found to vary with the accompanying choices. One notorious example in cognitive psychology is the speed-accuracy tradeoff (Luce, 1986; Pachella, 1974; Ratcliff, 1978; Wickelgren, 1977), a frequently observed phenomenon that a decision maker sacrifices accuracy for faster responses. This phenomenon had been a major challenge for cognitive psychologits because they couldn't conclude participants' performances or task difficulty based on how fast they responded or how many times they made errors. Instead, one has to examine both RT and responses simultaneously under the same, integrated framework (Ratcliff, Van Zandt, & McKoon, 1999), but it's not an easy task either, because RT is a continuous variable and skewed, and response is a categorical variable, with the error rate significantly lower than the rate of correct responses; few statistcal models could deal with such contrasting dependent variables in the same framework.

It is, therefore, crucial to study RT with great care regarding the points made above. In other words, first, one should examine the whole shapes of the RT distributions, not just mean or median, to compare RTs across different experimental conditions (Heathcote, Popiel, & Mewhort, 1991; Rouder, Lu, Speckman, Sun, & Jiang, 2005); second, one should consider the models of RT that can account for various sources of variability, with each source possibly linked to each of the substantive cognitive processes that govern the observed behavior (Schwarz, 2001; Vandekerckhove et al., 2010); and third, a successful model should be able to handle choice data as well as RT in the same theoretical framework. With these points in mind, I choose the diffusion model (Ratcliff, 1978) for the RT analysis. The diffusion model theory not only meets the three points listed above, but it also has been applied to numerous different fields with great success, making substantial implications about cognitive systems (Ratcliff, Gomez, & McKoon, 2004; Ratcliff & McKoon, 2008; Ratcliff, Smith, Brown, & McKoon, 2016). In the following sections, I discuss in detail the diffusion model.

3.1 Diffusion model

The diffusion model (Ratcliff, 1978) is one of the most widely applied cognitive models for two-choice tasks. It has been applied to lexical decision (Ratcliff, Gomez, & McKoon, 2004; Ratcliff, Thapar, Gomez, & McKoon, 2004; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008), semantic categorization task (Vandekerckhove et al., 2010), signal detection (Ratcliff & Rouder, 1998), numeracy judgments (Ratcliff, Thapar, & McKoon, 2010), perceptual tasks, such as brightness discrimination (Ratcliff, 2002; Ratcliff, Thapar, & Mckoon, 2003), motion discrimination (Palmer, Huk, & Shadlen, 2005), and orientation discrimination (Smith & Ratcliff, 2009). The theory has also been employed to examine deficits in neurological or psychological functions, such as aphasia (Ratcliff, Perea, Colangelo, & Buchanan, 2004), dyslexia (Zeguers et al., 2011), attention-deficit/hyperactivity disorder (ADHD) (Mulder et al., 2010; Weigard & Huang-Pollock, 2014), and schizophrenia (Moustafa et al., 2015). One of the reasons the model has been so popular is that it can account for RT distributions for both correct and error responses. The model also provides an account of the underlying dynamics of decision-making, which has helped deepen our understanding of cognition. Note that Ratcliff's diffusion model has achieved all these successes via the sequential sampling framwork. The sequential sampling process is so vital for the diffusion model, so we now turn our attention to the sequential sampling models.

3.1.1 Sequential sampling framework

Sequential sampling models, as the name suggests, use a conceptual idea of sequentially sampled information to describe the underlying decision-making process. Specifically, sequential sampling models assume that we accumulate samples of information about the given stimuli, and when the accumulated information meets a certain criterion, we make a decision. This general framework of the model is inspired by how our nervous system works (Luce, 1986). Our nervous system processes information about the external world, and the way it processes the information is not an all-or-none process that would happen at once, but rather an ongoing process that accumulates samples of information about the stimuli over time (Gold & Shadlen, 2007; Heekeren, Marrett, & Ungerleider, 2008). The information that enters the accumulation process are assumed to be noisy, which resembles the way our brain operates. For example, a sensory-related neuron would respond to a sensory stimulus (e.g., light, sound, etc.), but its output (usually measured in spikes per second) will vary even when a relevant stimulus is present though it tends to be higher (Gold & Shadlen, 2001). In the same way, sequential sampling models assume that each piece of sampled information is noisy, that is, at each time point, the sampled information may favor the wrong response; but in general, the information tends to be in favor of the correct response, eventually driving the decision-making process in the correct way.

One of the strengths of the sequential sampling models is that they provide theoretically coherent ways to account for accuracy and RT data, simultaneously. In cognitive psychology, the shapes of RT distributions are known to vary with accuracy (Luce, 1986; Pachella, 1974). Thus, if one is interested in studying RT, it is strongly advised to study the accompanying choices as well to get a deeper insight into our cognitive system (Luce, 1986; Ratcliff & Smith, 2004). But, this has been a challenge for many cognitive models. For example, signal detection theory models (Green & Swets, 1966) are considered one of the most influential theories in cognitive psychology, but the theory provides an account only of accuracy, not able to explain interactions of accuracy and RT (Ratcliff & Smith, 2004). The sequential sampling models don't have such limitations; indeed, this is where the sequential sampling models excel. Sequential sampling models not only provide a theoretical account of RT and accuracy within the same model, but they can also describe behavioral data (e.g., observed RT distributions) exceptionally well (Ratcliff et al., 2016; Vandekerckhove

& Tuerlinckx, 2007). The sequential sampling models have even been able to provide intuitive ways to account for non-trivial relationships between accuracy and RT; for instance, responses are in general slower and less accurate for difficult stimuli than those for less difficult stimuli (Luce, 1986; Pachella, 1974). Making such accounts of RT and accuracy, however, requires a detailed understanding of the behavior of the model, which I address below.

The core system of sequential sampling models again relies on information accumulation. The information to be accumulated is assumed to be noisy, so there are chances that the model can make wrong decisions. Such noisy information is sampled by a conceptual unit of the model, often called *accumulator*, until the model decides that enough information has been collected for a decision. In other words, sequential sampling models set a threshold for the amount of information, or *decision criteria*, and when the accumulated information hits one of the decision criteria, the model produces a decision associated with the criterion that is attained. In sum, sequential sampling models rely upon this core dynamic of the quality of information (i.e., how noisy the information is) and the quantity of information required for a decision (i.e., how much information is needed to make a decision). Firgure visualizes how these two components interact to generate a decision. Varying these two dimensions of the model, one is able to account for various relationships between accuracy and RT, many of which had been considered challenges in cognitive psychology.

Consider, for example, the speed-accuracy tradeoff (Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010; Wickelgren, 1977). This is a well-known phenomenon in cognitive psychology, where one sacrifices accuracy for faster responses, or vice versa. Although it may seem like an intuitive account of a relationship between RT and accuracy, it's not easy to analyze in practice. Imagine that we have a set of RT data and respective choices, recorded as correct or wrong. In this case, the first challenge we face would be simply detect the presence of the speed-accuracy tradeoff in the data.

Not every fast response sacrifices its accuracy; not every slow response is associated with improved accuracy either. One must dissociate the pattern of the speed-accuracy tradeoff from just slow or fast response patterns. Even if the researcher manages to spot a pattern of the speed-accuracy tradeoff in the data, he or she should be able to account for this phenomenon within the employed model, that is, using the parameters in the model. Sequential sampling models overcome those challenges via changes in the values of the decision criteria. For example, when in need of fast responses, a decision maker would lower the amount of information needed for a decision, so that he or she can make a decision fast; but at the same time, there are more chances that the noisy part of the information could lead to the wrong response because of the lowered criteria. When in need of accurate responses, on the other hand, a decision maker would raise the amount of information needed for a decision, so that he or she can accumulate as much information as possible. In this case, noisy information can hardly affect the decision process toward the wrong response, because a few wrong turns of sampled information could easily be made up by other pieces of information if there is enough time.

This core framework, represented as the information accumulation, is shared by all sequential sampling models, but different models assume different details about the accumulation process. For example, random walk models (Laming, 1968; Stone, 1960), one of the first sequential sampling models, assume that the information accumulation would happen at discrete time points, with the information being relative. That is, the information in favor of one alternative is automatically against the other alternative. If we instead assume that the information accumulates in an absolute way (i.e., the information for one alternative has nothing to do with the other alternative), a different type of models emerges: recruitment model (LaBerge, 1962), one of the first models of this kind, or more recently, race models (Brown & Heathcote, 2008; Marley & Colonius, 1992; Rouder, Province, Morey, Gomez, & Heathcote, 2015; Townsend & Ashby, 1983). All these models assume a similar accumulation process: For each of the given alternatives, there exists a counter that accumulates only the information in favor of the corresponding alternative. The accumulation process will stop whenever one of the counters reaches its criterion, and the first one will get chosen. Other than these models, sequential sampling models also include the models based on the recent findings in neurophysiology, for instance, the leaky competing accumulator model (Bogacz, Usher, Zhang, & McClelland, 2007; Usher & McClelland, 2001) and cortical network models (Amit & Brunel, 1997; Wang, 2002), and many other models as we vary certain aspects of the accumulation process (see Luce, 1986; Ratcliff & Smith, 2004, for a review).

The diffusion model (Ratcliff, 1978) assumes the accumulated information to be relative and continuous, with the decision criteria being stable over time. In other words, the model assumes only one accumulator that accumulates information until either one of decision criteria is met. When compared to the aforementioned complicated models, like neurophysically inspired models, such as the cortical network model (Wang, 2002), the diffusion model seems to offer only a simple mechanism of decision-making process. However, I argue that such simplicity of the diffusion model is an advantage, rather than a limit for the following reasons. First, due to its simplicity and straightforward process, the diffusion model has been widely applied to numerous fields, while not sacrificing the ability to grasp the essence of our cognitive system. Indeed, the diffusion model is considered one of the most influential models in lexical decision task and recognition memory search task, which demonstrates that the diffusion model is well capable of providing sound accounts of cognition.

Second, despite being a sequential sampling model, the diffusion model is relatively easy to apply to data. Generally, sequential sampling models are extremely challenging to run on data due to its stochastic processes (Schwarz, 2001). The diffusion model itself, for example, had not been considered a popular option for data analysis until very recently because of its infinite oscillating series in the expression for the cumulative density function, which, depending on the parameteric assumptions, possibly leads to intractable integrals (Vandekerckhove & Tuerlinckx, 2007; Ratcliff & Tuerlinckx, 2002). But, with recent development in statistical techniques, running diffusion models nowadays is just a statistics package away. There have been various packages, or softwares developed specifically for diffusion models as of this writing (e.g., fast-dm, Voss & Voss, 2007; DMAT, Vandekerckhove & Tuerlinckx, 2007; HDDM, Wiecki, Sofer, & Frank, 2013), and most notably, Stan (Carpenter et al., 2017; Stan Development Team, 2022), a general statistical modeling platform, includes the diffusion model as one of its default distributions. Without those packages, only few applied researchers could ever apply the diffusion model, not to mention the above neurophysically inspired models, to their data.

Finally, parameters in diffusion models have shown to be reliably estimated. For example, in Lerche and Voss's (2017) lexical decision study, the main parameters (i.e., drift rate and decision criteria) of the diffusion model show high test-retest reliability, with all of its correlations, rs, greater than .7. Also, in another study of lexical decision task (Yap, Sibley, Balota, Ratcliff, & Rueckl, 2015), the researchers found that the key parameters of the diffusion model all show moderate-to-high between-session reliability (rs from .39 to .72). Given that the parameters of diffusion models are often used as a diagnostic tool (e.g., Aschenbrenner, Balota, Gordon, Ratcliff, & Morris, 2016), it is critical to have such good reliability in estimating parameters. However, when it comes to parameter recovery simulations, diffusion models have shown mixed results. When the model is set to its full complexity, that is, all of the main parameters (except for the decision criteria) of the diffusion model are allowed to vary across trials, the model often fails the parameter recovery simulations (Lerche & Voss, 2016; Ratcliff & Tuerlinckx, 2002; van Ravenzwaaij & Oberauer, 2009). It is the across-trial variability that has hard time getting recovered in simulations. So, a recommended practice of fitting the model is to limit the model's complexity by setting some, or all, of the across-trial variability to be 0 (Boehm et al., 2018), but this may not be feasible depending on the goal of the study, or experimental design. Hence, it is important to check the ability of the model in use to recover the true parameters prior to application to real data.

In this section, I've discussed the core structure of diffusion models, the sequential sampling framework, and why this class of models has received so much attention in cognitive psychology and, more recently, in neurophysiology. In the next section, I examine diffusion models in detail, with a focus on its parameters and mathematical model.

3.1.2 Parameters of the diffusion model

The standard diffusion model has four main parameters and three across-trial variability parameters: drift rate, decision criterion (sometimes called boundary separation, or upper threshold), starting point (or bias), and non-decision time (see Figure 3.1) are main parameters, and across-trial variability in drift rate, starting point, and non-decision time are the parameters that governs how much the main parameters vary from trial to trial. As one of the sequential models, the diffusion model relis on the information accumulation process to generate a decision. This process is assumed to occur through a one-dimensional internal representation of the *accumulator*, a conceptual unit that accumulates the information about the given stimuli once the stimuli are presented. The main parameters, except the non-decision time, of the diffusion model regulate the behavior of the accumulation process; specifically, those parameters govern where the accumulation process begins at (starting point), how fast the accumulation process would reach the threshold and make a decision (drift rate), and how much of information must be accumulated for a decision (decision criteria). Then, the non-decision time is added to the time taken for this informa-

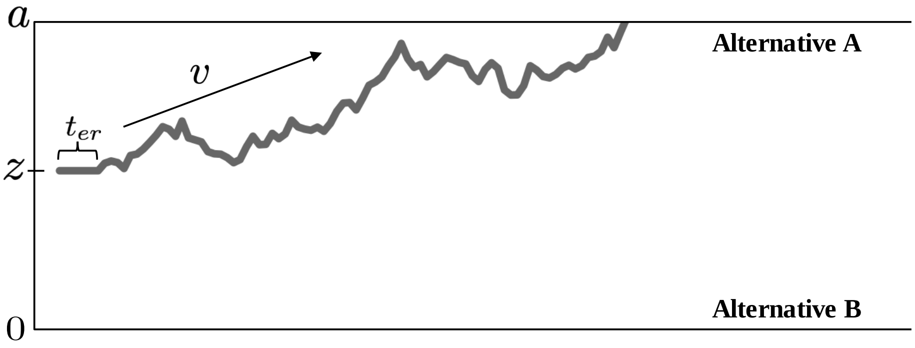


Figure 3.1: A basic framework of the diffusion model. a represents the decision criterion, z represents the starting point of the accumulation process, v represents the drift rate, and t_{er} represents the non-decision time. When the accumulator reaches one of decision criteria (either 0 or a), the decision is made.

tion accumulation process, completing the model's prediction for RT. In the next few paragraphs, I will give a brief review of each of these parameters to provide a better understanding of the model.

The first parameter of the diffusion model is the drift rate, denoted by v for a mathematical formulation of the model. The drift rate is the mean rate of information uptake, representing the quality of information about the given stimuli. That is, the actual amount of information that enters the model at every time point is not deterministic, but variable, or noisy. Such noisy information is due in part to the assumption about our sensory systems that the information encoded by our sensory systems is probabilistic (Green & Swets, 1966) and, more recently, supported by the pattern of neuronal firing, which has been found to vary in its output even when a related stimulus is present (Gold & Shadlen, 2001). So, the diffusion model treats the information a random variable, with its distribution being a normal distribution with mean of v (i.e., drift rate) and variance of s^2. In other words, the drift rate determines the speed of information processing; the greater the drift rate, the faster the response, and the more likely the model is to predict correct responses. The standard deviation s governs the degree of variation of information over time. This parameter typically serves as a scale parameter, meaning that it provides a scale upon which the parameters of the diffusion model are all estimated. Thus, the standard deviation is fixed to any arbitrary number; generally, it is set to be 0.1, which is common in most studies (Vandekerckhove & Tuerlinckx, 2007), but a different value has also been used (Navarro & Fuss, 2009). If one decides to add across-trial variability to drift rate, denoted by s_v, the drift rate on each trial, which is now denoted by ξ, follows normal distribution with mean of v and standard deviation of s_v. Note that when s_v is fixed at 0, the same drift rate applies for every trial, so ξ becomes v.

The second parameter is the decision criterion, denoted by a. The decision criterion determines the amount of information needed for a decision. In other words, this parameter determines when to stop the accumulation process, and so, it is directly related to the account of the speed-accuracy tradeoff. If the parameter is set to a large value, more information is needed to make a decision. Thus, RT is slow and the choice is accurate because the decision is made based on the large amount of information. If the parameter is set to a small value, less information is needed to make a decision. As a result, RT is generally fast and the choice is less accurate because there are more chances that the noisy part of each sampled information drives the accumulation process to the wrong criterion (Bogacz et al., 2010). Many empirical results have shown that this parameter can be directly manipulated by experimental instructions. For example, when the experimenter emphasizes speedy responses (e.g., "Try to make responses as fast as possible"), participants' decision criteria are generally estimated to be low; when the experimenter emphasizes accurate responses (e.g., "Try to make responses as accurate as possible"), participants' decision criteria are generally estimated to be high (Voss, Rothermund, & Voss, 2004; Wagenmakers, Ratcliff, et al., 2008). These results show that participants can adapt their decision-making strategies to the context where the decision is made, particularly by making a tradeoff between RT and accuracy.

The third parameter is the starting point, denoted by z. The starting point is often called a bias, because it determines where the accumulator starts its accumulation process at, even before the process itself begins. The starting point, like the parameters discussed above, affects the probability of choices and RT; if the accumulation process begins near one of the decision criteria, the process is more likely to end up at the closer criterion, with faster RT. In most cases, it is reasonable to assume unbiased responses (e.g., Ratcliff, Thapar, & McKoon, 2001), so when in need of a simplified model, it's this parameter that researchers would first consider setting fixed for the purpose of simplification (Wagenmakers, Van Der Maas, & Grasman, 2007). One can assume across-trial variability in starting point s_z. This is because when it comes to describing RT distributions for errors, variability in starting point is a necessary component for the model (Laming, 1968; Ratcliff & Rouder, 2000). In this case, on every trial, the starting point is assumed to be drawn from uniform distribution with mean of z and range of s_z.

The last parameter is the non-decision time, denoted by t_{er}. This parameter accounts for the time taken for all extra processes for decision-making other than the information accumulation process. Two prominent examples of this type of processes are stimulus encoding and motor response (e.g., press the right key on keyboard, or move the mouse cursor and click on it). Since every person has a different speed of encoding of the stimulus and motor response, this parameter is considered a random variable, usually assumed to follow a uniform distribution. Then, the diffusion model predicts the observed RT as the sum of the time taken for the accumulation process, denoted by t_d, and non-decision time t_{er}. When the across-trial variability is assumed for non-decision time $s_{t_{er}}$, non-decision time for each trial is assumed to be drawn from uniform distribution, like the starting point mentioned above, with mean of t_{er} and range of $s_{t_{er}}$.

Deciding whether to include across-trial variability to the parameters typically depends on the goal of the study or the type of data. For example, if we allow the drift rate to vary every trial, the model can account for slower RT for errors than for correct responses (Ratcliff, 1978). If we allow the starting point to vary every trial, the model can account for faster RT for errors than for correct responses (Laming, 1968). If we combine these two types of variability, the diffusion model is able to account for a mixture of those two effects, that is, errors are slower than correct responses for difficult tasks (i.e., accuracy is low) and errors are faster than correct responses for less difficult tasks (i.e., accuracy is high) (Ratcliff & Rouder, 1998, 2000; Ratcliff et al., 1999). Note, however, that such flexibility of the model comes at a price. As mentioned above, fully extended diffusion models often fails the parameter recovery simulations (Boehm et al., 2018; van Ravenzwaaij & Oberauer, 2009) and is not even feasible to fit when the data are not enough (Wagenmakers et al., 2007). Thus, it is important for researchers to be aware of their models' limits and advantages with regard to the setting where the models are applied, which particularly inspires the parameter-recovery simulation study in the next chapter.

3.1.3 Mathematical formulation of the diffusion model

The probability density function (PDF) for the diffusion model, especially with across-trial variability in main parameters, easily involves intractable integrals because of its infinite series, so estimating the parameters often requires some kinds of approximation methods. In this dissertation, I employ the particular approximation method suggested by Navarro and Fuss (2009). This method is concerned only with the Wiener diffusion process, not with the full diffusion model, which describes the information accumulation process, where the information accrues over time until it crosses one of the decision criteria. While the Wiener diffusion process accounts for the core structure of the diffusion model (Ratcliff, 1978), it doesn't account for across-trial variability parameters. As the authors argued, a huge benefit of this approach lies in its quick and accurate calculation of the PDF for the Wiener process, and this is due largely to the use of two different approximation methods for "small" time expansion and "large" time expansion. I briefly review its basic ideas in this section. For interested readers, I strongly recommend to consult the paper (Navarro & Fuss, 2009) directly.

The Wiener process can be derived from the random walk model, a discrete-time stochastic model, as its limiting case. Suppose $X(t)$ is a real-valued random variable, representing the state of information at time t. If we make t infinitely small, the system can be expressed via the stochastic differential equation $\frac{d}{dt}X(t) \sim \text{Normal}(v, s^2)$. If we assume that the initial state of information $X(0)$ lies in a range of $0 < X(0) < a$, and that a decision is made at the first time t for which $X(t)$ is greater than or equal to a (i.e., $X(t)$ hits the upper boundary), or smaller than or equal to 0 (i.e., $X(t)$ hits the lower boundary), then it results in the distribution of choice, c, and RT, t_d, referred to as the Wiener first-passage time (abbreviated to Wiener) distribution. If we fix the variance of the Wiener process, s^2, at 1, as s serves as a scale parameter here, the behavior of the Wiener distribution is governed by three parameters: drift rate (v), decision criterion, or upper boundary (a), and starting point (z). Therefore, if choice and RT are both random variables and follow the Wiener distribution, we write it as $(c, t_d) \sim \text{Wiener}(v, a, z)$.

The probability density function for the Wiener distribution, $f(t|v, a, z)$, is given by Feller (1968), which provides the probability of the diffusion process reaching the lower boundary at time t. In expressing its PDF, Navarro and Fuss (2009) chose to use the relative starting point $w = z/a$, which varies from 0 to 1, instead of using the absolute starting point, z, for mathematical convenience. With this new parameterization, The PDF for the Wiener distribution, $f(t|v, a, w)$, is given by:

$$f(t|v, a, w) = \frac{\pi}{a^2}\text{exp}\left(-vaw - \frac{v^2t}{2}\right)\sum_{k=1}^{\infty} k\text{exp}\left(-\frac{k^2\pi^2 t}{2a^2}\right)\sin(k\pi w).$$

If we are interested in the probability density at the upper boundary at time t, we can simply substitute v with $v' = -v$ and w with $w' = 1 - w$. One obvious difficulty involved in this PDF is evaluating the infinite series. One way to deal with such infinite series is to set the threshold and stop calculating terms once the sum exceeds the pre-determined threshold value. But, the problem is, the PDF of the Wiener distribution behaves differently when t is small and when t is large. So, the authors first analytically derived a solution for when the change happens (i.e., solved for the exact t for which t is considered small or large), then applied different approximation methods to "small time" and "large time" representations of the Wiener densities. The resulting distribution shows effective performance, measured by time taken to achieve a desired accuracy, in multiple simulation studies, compared with the method suggested by Ratcliff and Tuerlinckx (2002).

The given approximation of the PDF has been implemented in the `RWiener` (Wabersich & Vandekerckhove, 2014) `R` package (R Core Team, 2022), which is again implemented in Stan (Stan Development Team, 2022) as one of its default PDFs. Since I use Stan as my primary statistical tool for all the Bayesian analyses conducted in this dissertation (except for the Bayes factor calculation for representational theories), the fact that I can easily access the PDF of the Wiener distribution within Stan provides me with massive advantage in conducting Bayesian analysis of the diffusion model. Now, we turn to the application of the diffusion model to preferential choice tasks, where I discuss issues regarding such applications and the format of the data to be analyzed.

3.2 Applying the diffusion model to preferential choice tasks

3.2.1 Background

In the past few decades, there has been an increasing interest in the underlying mechanism behind observed choices in economics. Such interest has always been around in economics, but with recent advances in technology, especially with neurophysiology, many economists started considering implementing the newly available technology (e.g., fMRI, EEG) in developing theories that address not just the outcomes, but its underlying processes as well (Camerer, 2013; Webb, Levy, Lazzaro, Rutledge, & Glimcher, 2019; Masatlioglu, Nakajima, & Ozbay, 2012). However, the benefits of such technological advancements come at a price; the endeavor to incorporate the brain activity into economic models requires specialties in neuroscience, which most economists lack (Camerer, 2013), and those models almost always necessitate the analysis of RT, which presents the same challenge psychologists had faced earlier: numerical and statistical difficulties (Spiliopoulos & Ortmann, 2018; Vandekerckhove & Tuerlinckx, 2007). Nevertheless, an increasing number of studies in economics pay more attention to process-based models than to outcome-based models, and those studies include RT as a meaningful dependent variable.

One notable example of this trend of research in economics is the increasing use of sequential sampling models. Fudenberg and his colleagues (2018), for example, used the seqeuntial sampling framework for their new model, the uncertain-difference drift diffusion model, which exploits the agent's prior beliefs about the utility difference between the given choice alternatives in constructing the joint distribution of choice probabilities and RT. Clithero (2018) showed that the diffusion model can improve the out-of-sample predictions for food choices, compared to the models that analyze the outcome data only, such as, logistic regressions. Also, Konovalov and Krajbich (2019) used the standard diffusion model (Ratcliff, 1978) to predict strength of preference, along with individual utility-function parameters. From these results, they were even able to predict which choices are likely to be reversed later.

However, the type of task typically administered in economics is not what the diffusion model had originally targeted. As Ratcliff warned in many papers (e.g., Ratcliff & McKoon, 2008; Ratcliff & Rouder, 1998; Ratcliff et al., 2016), the diffusion model is not supposed to be used for value-based decisions, like preferential choice tasks (e.g., food choices, or gamble choices), but only for speeded decision-making tasks, like perceptual decisions (e.g., brightness discrimination task). The reason lies in the different decision-making processes. The process of most speeded decisions rely on the information about the given stimuli. Determining which dot is brighter, for example, obviously requires one to collect information about the stimuli on screen. The same is true for motion discrimination task, color discrimination task, or any other tasks typically administered in cognitive psychology. For those tasks, the diffusion model makes sense, because the likelihood of the correct answers depends on how much information about the stimuli has been processed, and such process can directly be representable as the information accumulation process of the diffusion model. But, value-based decisions require different types of information. The choice between banana and apple doesn't need a decision maker to take a close look at the stimuli. Most people are already familiar with what a banana or an apple looks like. Instead, they might need to consider nutrient information of these two fruits, or think of some memories associated with each of them. Thus, the value-based decisions are considered one of "the multiple-stage processes that might be involved in, for example, reasoning tasks (Ratcliff & McKoon, 2008, p. 875)," and not recommended for the diffusion model analysis.

Recent findings in neurophysiology, however, showed different evidence for value-based decisions. For example, Krajbich and his colleagues (2015) successfully applied a sequential sampling model to a food choice task and were even able to predict behaviors in social-decision tasks using the sequential sampling model. Basten and his colleagues (2010) found empirical evidence that our brain integrates the information about costs and benefits when making value-based decisions, in the same way as we respond to perceptual decision making. Also, Polanía and his colleagues (2014) analyzed EEG recordings of value-based decisions and perceptual decisions and found that there were common areas in the brain activated for both types of tasks, and this area was related to the information accumulation process. Such findings in neurophysiology all make the same implication about the underlying mechanism: our brain makes value-based decisions in the same way as perceptual decisions. Although different types of decisions may require different types of information to be accumulated, our brain would go through the same information accumulation process for both types of decisions (Milosavljevic, Malmaud, Huth, Koch, & Rangel, 2010; Shadlen & Shohamy, 2016). And it is such neurophysiological findings that have particularly motivated me to apply the diffusion model to preferential choice tasks.

3.2.2 The data to be examined

The experimental paradigm of interest in the current analysis is the preferential choice task, the same paradigm Tversky (1969) employed for his own experiment. In an experimental session, a series of pairs of monetary gambles appear one at a time, and participants are asked to express their preferences between the given pair of gambles on each trial. Then, the data consist of which gamble is chosen on each trial, along with the time taken to make that choice, per individual. The number of gambles for the current analysis is five, i.e., $A,\ B, ...,\ E$, so the number of unique gamble pairs that the five gambles make is $\binom{5}{2} = 10$. One unique feature about the preferential choice task paradigm is the repetition of the same gamble pairs. For example, the gamble pair A and B is designed to appear multiple times in the same experimental session.

This is due to Tversky's (1969) observation that our choices are not deterministic, but rather variable even when the same task is given repeatedly. Thus, in order to grasp the probabilistic nature of our choices, it is common for preferential choice experiments to include the same gamble pairs multiple times. One's choice variability is then measured by a choice proportion for each gamble pair. Traditional approaches for analyzing data are usually centered on such choice proportions, which are the maximum likelihood estimates of choice probabilities (Myung, 2003). But, now with the diffusion model, we utilize both choice data and RT to examine the underlying cognitive processes, assuming each trial is a whole diffusion process.

In the following sections, I give an overview of the particular version of the diffusion model I choose to apply to the preferential choice task paradigm. This involves accounts of parametric assumptions for the given data and priors for the main parameters of the diffusion model in order to perform the Bayesian analysis.

3.3 Model specification

3.3.1 Parameter constraints

The diffusion model varies its form depending on the constraints on the parameters of the model. If the across-trial variability is not allowed and the starting point is fixed at the middle point, it will result in the model, known as the EZ-diffusion model (Wagenmakers et al., 2007). If the EZ-diffusion model is too restrictive for the data, one can even extend the EZ-diffusion model to include the starting point (Grasman, Wagenmakers, & van der Maas, 2009; Wagenmakers, van der Maas, Dolan, & Grasman, 2008). Or, one can set only some of across-trial variability to 0 and allow others to be estimated from data. The choice of the model is often considered a tradeoff between its versatility and robustness, so it depends on the goal of the study and the data at hand.

The model I choose is the hierarchical diffusion model, with its core model being the Wiener distribution (Navarro & Fuss, 2009), and any additional parametric assumptions added via the hierarchical structure. This is the exact framework that Vandekerckhove and his colleagues (2011) recommended for the diffusion model analysis. The strength of this way of analysis is that we can take advantage of the hierarchical structure to remedy some of the problems the diffusion model has long been suffered (Rouder et al., 2005; Vandekerckhove et al., 2010). In hierarchical models, all individual-level parameters are assumed to be drawn from group-level distributions. So, when individual parameters are estimated from data, the hierarchical structure allows the group-level parameters to be estimated from the data at the same time. By doing so, the estimation of any particular individual parameter can be informed by other individuals' parameters, through the group-level distributions. The benefits of the hierarchical approach are: first, the estimation process is less vulnerable to extreme outliers, and second, the parameters are reliably estimated even when the number of trials per experimental condition is small, or not equal between the conditions (McElreath, 2020).

Specifying the hierarchical model starts with the model for observed responses. In this specification, I generally follow the mathematical notation of Vandekerckhove et al.'s (2011). Suppose p indexes people, $p \in \{1, ..., P\}$, i indexes conditions, $i \in \{1, ..., I\}$, and j indexes trials, $j \in \{1, ..., J\}$. If we write the observed responses as a vector $\boldsymbol{y}$ that consists of observed choice c and RT t, the response of person p in condition i on trial j can be written as $\boldsymbol{y}_{(pij)} = \begin{pmatrix} c_{(pij)} \\ t_{(pij)} \end{pmatrix}$. Then, $\boldsymbol{y}_{(pij)}$ is distributed according to the the Wiener distribution:

$$\begin{pmatrix} c_{(pij)} \\ t_{(pij)} \end{pmatrix} = \boldsymbol{y}_{(pij)} \sim \text{Wiener}(a_{(pij)}, z_{(pij)}, t_{er(pij)}, v_{(pij)}),$$

where $a_{(pij)}$ is the decision criterion (or boundary separation), $z_{(pij)}$ are starting points, $t_{er(pij)}$ are non-decision time, and $v_{(pij)}$ is the drift rate. Note that the current model has all the parameters accompanied by indices pij placed between parentheses. As in Vandekerckhove et al., running indices are put in parentheses to avoid confusions with other types of subscripts; for example, the non-decision time parameter t_{er} has subscript er, but it is not surrounded by parentheses, so it's a part of the parameter's name, not a running index. The parameters with (pij) indicate that they are allowed to vary across people, conditions, and trials. If one wants to know whether the model allows for across-trial variability in parameters, he or she could simply investigate the parameters to see whether they have the running index (j). This is the basic setup for the model for responses, and now I will put constraints on the parameters as I proceed.

Experimental condition The first constraint I impose on the model is the experimental conditions. It's rather a choice of the experimental design, than a constraint, where I choose not to include any experimental conditions. My desire for this decision is that the model can be made as parsimonious as possible by not assuming different conditions for better performance on predicting the main parameters. Hence, I drop index i from all parameters.

Starting point The starting point is known to be a crucial parameter for describing the RT distributions for errors (Ratcliff & Rouder, 1998). However, the present study runs the diffusion model analysis on preferential choice data, where participants are asked to choose between two monetary gambles. In this case, it doesn't make sense to regard either choice as error, because both alternatives can be chosen with different preferences. So, any properties of RT distributions for errors are not pursued in the current analysis. Therefore, I decided to forgo the estimation of the starting point, that is, the starting point is fixed at 0.5, $z_{(pj)} = 0.5$, for all participants.

Across-trial variability When it comes to across-trial variability, many studies suggested that the inclusion of the across-trial variability didn't increase the power of estimating the main parameters (i.e., drift rate and decision criterion) (Lerche & Voss, 2016). Thus, I decided not to estimate across-trial variability for all parameters. One exception is the drift rate. In preferential choice tasks, drift rate is usually determined by the difference in subjective value (sometimes referred to as utility) between the given alternatives (Fudenberg et al., 2018; Polanía et al., 2014; Webb et al., 2019). For example, if the gamble pair A and B is present on trial j, decision maker p would determine drift rate v_{pj} by computing the value difference between A and B, that is, $v_{(pj)} = u_{A(p)} - u_{B(p)}$, where $u_{A(p)}$ denotes the subjective value of Gamble A for person p. Thus, the estimation of drift rates is equivalent to estimating subjective values of the gambles.

Final model The final model with all the constraints listed above is as follows:

$$\boldsymbol{y}_{(pj)} = \begin{pmatrix} c_{(pj)} \\ t_{(pj)} \end{pmatrix} \sim \text{Wiener}(a_{(p)}, z_{(pj)} = 0.5, t_{er(p)}, v_{(pj)}), \tag{3.1}$$

where $v_{(pj)}$ is determined by value difference between the given gambles on trial j. Note that the drift rate $v_{(pj)}$ doesn't vary every trial. Preferential choice tasks typically include the same gamble pair multiple times, so the same values of drift rate will appear every once in a while during an experimental session. For statistical inferences on the parameters, I use a Bayesian hierarchical model as mentioned above, where priors of the parameters also have their own priors, creating the hierarchy. Thus, it is necessary to provide priors for all the parameters in the model, including the priors of priors (i.e., hyperpriors), in order to get the Bayesian analysis to work. I specify priors in detail in the following section.

3.3.2 Priors for parameters

The main model I use for the diffusion model analysis is specified in Equation (3.1), where no across-trial variability is allowed for any parameters, and starting point z is fixed at 0.5, indicating an unbiased process. One unique feature of the model is the way it computes drift rates. Recall that the data for analysis are preferential choices between pairs of monetary gambles. Drift rates are determined by the difference in subjective value between the given gamble pair. This way of computing drift rates is consistent with literature especially when the diffusion model is used for value-based decisions (Fudenberg et al., 2018; Milosavljevic et al., 2010; Polanía et al., 2014). Thus, instead of setting prior for drift rate directly, we need to set prior for the subjective value of each gamble, then drift rates are computed via the estimated subjective values.

For Bayesian analysis, all the parameters in the model require priors to be specified. In Equation (3.1), the parameters that need priors are $\boldsymbol{\theta}_{(p)} = (a_{(p)},\ t_{er(p)},\ v_{(pj)})$, where $v_{(pj)}$ are computed via subjective values $\boldsymbol{u}_{(p)} = (u_{A(p)}, ..., u_{E(p)})$. So, priors need to be set for $\boldsymbol{u}_{(p)}$, instead of $v_{(pj)}$. I begin with priors for $a_{(p)}$.

Decision criterion a The decision criterion parameter a governs the amount of information to be accumulated to produce a decision. It is via this parameter that the diffusion model accounts for the speed-accuracy tradeoff (Bogacz et al., 2010). Also, a determines the style of decision making, either conservative, i.e., more information is needed for a decision, or less-conservative, i.e., less information is needed for a decision (e.g., Ratcliff, Thapar, Gomez, & McKoon, 2004). From the perspective of the representational theories of preferences, a is exactly the parameter that should reflect the different cognitive processes used by transitivity and lexicographic semiorders. These two theories uses clearly different cognitive processes, where transitivity, as a core component of all rational economic theories, characterized its decision-making process by exploiting all relevant pieces of information; lexicographic semiorders, on the other hand, inspire numerous heuristic-based decision-making strategies, which are typically represented by the less-is-more principle, that is, ignoring information. Since the main difference between the two theories is the amount of information that goes into decisions, such different cognitive processes should manifest through the decision criterion parameter a, because it governs the amount of information needed for a decision.

Suppose that we have the classification data, where participants are classified into either transitivity or lexicographic semiorders. Let $X_a = (X_{a(1)}, ..., X_{a(P)})^T$ be a vector of size P taking values of either 1 or 0: if participant p is classified to transitivity, $X_{a(p)} = 1$; otherwise, $X_{a(p)} = 0$. The vector X_a enters the model like an explanatory variable in the regression model. Let β_{0a} and β_{1a} be the intercept and coefficient of X_a, respectively. Then, the prior for $a_{(p)}$ is,

$$log(a_{(p)}) \sim \text{Normal}(\mu_{a(p)},\ \sigma_a^2),\ \ p \in \{1, ..., P\}$$

$$\mu_{a(p)} = \beta_{0a} + \beta_{1a} X_{a(p)}.$$

Note that a is log-transformed here, and the rest of the model for a is defined on a log scale. This is because a must be positive in this model, and the log-transformation ensures its value to be positive.

The prior for decision criterion a takes the form of regression model, where $X_{a(p)}$

enters the model as a dummy variable. In this parameterization, β_{0a} behaves like an intercept, accounting for the baseline level of a for all participants, while β_{1a} accounts for the additional effect that transitivity has on a, just like any other regression models. That is, for those classified to lexicographic semiorders, mu_a comes down to just β_{0a}, since $X_{a(p)}$ equals 0 for the lexicographic semiorder classification; for those classified to transitivity, mu_a becomes $\beta_{0a} + \beta_{1a}$. So, we can test whether β_{1a} equals 0 to see if there is any group difference in a, serving as a main hypothesis testing tool. In Chapters 4 and 5, I exploit this way of testing the hypothesis against simulated and real data sets.

Note that the above prior has another set of parameters ($\beta_{0(a)},\ \beta_{1(a)}$), which need its own priors as well. This setup is usually called a hierarchical, or sometimes multilevel, Bayesian model, where the hierarchy arises in the structure of priors. In this setting, the priors of priors are referred to as hyperpriors (Gelman et al., 2013). I employ the following normal distributions for the hyperpriors of ($\beta_{0(a)},\ \beta_{1(a)}$):

$$\beta_{0a} \sim \text{Normal}(\nu,\ \omega_1^2),$$
$$\beta_{1a} \sim \text{Normal}(0,\ \omega_2^2),$$

I consider normal distributions for hyperpriors of $\beta_{0a},\ \beta_{1a}$, because when no information other than its mean and variance (or that, at least, the variance is finite) is available, normal distributions are the safest bet for a distribution from the perspective of maximum entropy (McElreath, 2020). In other words, if we have a group of continuous random variables with finite variance, and with no other constraints imposed, normal distributions are the most likely distribution that will naturally emerge for these random variables. That is, numerous different processes in nature would end up producing outcomes that can be approximated by normal distributions. This is because normal distribution is the one that spreads out its probability density most evenly over its domain under the constraint of finite variance (see McElreath, 2020, for more discussion on this topic). Moreover, I use Stan (Carpenter et al., 2017; Stan Development Team, 2022) to make Bayesian inferences in this study, and Stan doesn't require conjugacy of priors thanks to its advanced sampling algorithm (it implements Hamiltonian Monte Carlo and its extension, the No-U-Turn sampler). Thus, there are no mathematical limits that keep us from employing normal distributions as hyperpriors.

For σ_a, I use exponential distribution for its prior:

$$\sigma_a \sim \mathrm{Exp}(\lambda).$$

The exponential distribution ensures its values to be positive, and most of its probability density is concentrated around 0, which I think is suitable for the prior of σ_a.

Drift rates v, or subjective values $u_A, ..., u_E$ As mentioned above, drift rates are determined by subjective value differences between the two given gambles. Let $v_{AB(p)}$ be the drift rate of participant p when the gamble pair A and B is given. Then, the drift rate $v_{AB(p)}$ is given by

$$v_{AB(p)} = u_{A(p)} - u_{B(p)},$$

where $u_{A(p)}$ and $u_{B(p)}$ are subjective values that participant p has on gambles A and B, respectively. In the above parameterization, we see that drift rates are just linear combinations of subjective values, so it's the subjective values that we need to estimate from data. Also, a close examination reveals that subjective values determine drift rate via subtraction, so only the difference between two subjective values matters; for example, $u_A = 500$ and $u_B = 480$ will result in the same drift rate as $u_A = 30$ and $u_B = 10$ will do. Thus, in order to make the model identifiable, I fix the subjective value of gamble C at 0, estimating other subjective values relative to gamble C. I consider the following prior for subjective values:

$$\begin{pmatrix} u_{A(p)} \\ u_{B(p)} \\ u_{D(p)} \\ u_{E(p)} \end{pmatrix} = \boldsymbol{u}_{(p)} \sim \text{MVN}_4 \left(\begin{pmatrix} \mu_A \\ \mu_B \\ \mu_D \\ \mu_E \end{pmatrix}, \Sigma \right), \quad p \in \{1, ..., P\}, \tag{3.2}$$

where $\text{MVN}_n(\boldsymbol{\mu}, \Sigma)$ is a n-dimensional multi-variate normal distribution with mean vector of size n, $\boldsymbol{\mu}$, and $n \times n$ variance-covariance matrix Σ.

As with the decision criterion parameter a, the prior specified in Equation (3.2) has parameters $(\mu_A, ..., \mu_E)^T$ and Σ that require their own priors to be specified. For mean $(\mu_A, ..., \mu_E)^T$ of the multi-variate normal distribution, I simply employ normal distribution for each component of the mean vector, with mean of 0 and the same variance, independently.

$$\mu_k \sim \text{Normal}(0, \xi^2), \quad k \in \{A, B, D, E\}.$$

When it comes to setting prior for Σ, however, we cannot employ any priors because Σ is a variance-covariance matrix, so it must meet a requirement that the matrix be non-negative definite. A typical strategy is to employ the inverse-Wishart prior for the variance-covariance matrix (Gelman & Hill, 2007), or to decompose the variance-covariance matrix into the diagonal matrix of standard deviations and correlation matrix, and set the priors in terms of standard deviations and correlation matrix (Barnard, McCulloch, & Meng, 2000). But, for the current model, I take a slightly different approach. When it comes to modeling subjective values of gambles, it is quite intuitive to suppose that similar gambles (e.g., Gambles A and B) would be associated with similar subjective values and dissimilar gambles (e.g., Gambles A and E) would be associated with different subjective values. In other words, the distance between two gambles can be a critical factor for determining subjective values, where the distance here means how far the given gambles are located within the chain (i.e., A, B, C, D, E); that is, the distance represents a measure of similarity in gambles' attributes (i.e., payoff and probability of winning). If the distance matters in determining subjective values, we should let the prior for Σ reflect this information. To this end, I employ the prior suggested by McElreath (2020) for Σ, where the prior considers the geographic distance between two objects for their covariance. Let Σ_{ij} be the covariance between any pair of gambles i and j. Then,

$$\Sigma_{ij} = \eta^2 \text{exp}(-\rho^2 D_{ij}^2) + \delta_{ij}\sigma^2, \quad i, j \in \{A, B, D, E\}, \tag{3.3}$$

$$\text{where } \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{3.4}$$

In Equation (3.3), the main term is $\text{exp}(-\rho^2 D_{ij}^2)$, where D_{ij}^2 is the squared distance between two objects i and j, and ρ^2 is the parameter that governs how fast or slowly the covariance declines as the distance between two objects grows larger. Then, η^2 determines the maximum covariance any two different objects can possibly have. That is, in order to achieve the maximum covariance between two objects in this formula, the distance between the two objects should be made as small as possible. As the distance approaches 0, the term $\text{exp}(-\rho^2 D_{ij}^2)$ will get close to 1, which result in the covariance close to η^2. Note that when it comes to variance, i.e., when $i = j$, the distance measure D_{ij} will be 0 because the distance of one object from itself is 0, so the above formula comes down to $\eta^2 + \sigma^2$. In that case, σ^2 serves to account for extra variance, beyond η^2 when $i = j$.

For this formula to work for Bayesian analysis, we need to feed the following information to the model: distance matrix D of gambles, and priors for $\eta, \rho,$ and σ. Since it's gambles, not physical objects we can measure the distance between, we need a way to compute the distance between gambles to build distance matrix D. As pointed out before, similar gambles are presumably associated with similar subjective values, so the distance should somehow be a measure of similarity between a pair of gambles. The question is, how should we determine similarity between a pair of gambles?

Recall that the monetary gambles have two attributes: payoff and probability of winning. Different gambles are defined by different values of these two attributes. If one considers that two gambles are similar, the two gambles should be similar in its attributes, that is, they would have similar payoffs and probabilities of winning. If we consider each attribute an axis of the coordinate system, each gamble can be located as a unique point, with its x-coordinate being its payoff and y-coordinate being probability of winning. And similar gambles will be located close to each other, whereas dissimilar gambles should be located far apart. Then, the distance between two points naturally represents the degree to which the two gambles are similar in this coordinate system, so I compute the distance between two gambles for a measure of similarity. The particular formula I employ to compute the distance is the Euclidean distance formula. This is perhaps the most common way to compute the distance between any two points in a coordinate system, but we need to make sure the two attributes are on the same scale prior to the computation of distance. So, I first standardize each attribute of gambles, using its mean and standard deviation, then compute the distance using the standardized values. For example, Gamble A from Set 1 is (\$25.43, $\frac{7}{24}$), but after standardization, it becomes $(1.265, -1.265)$. With these standardized values, the Euclidean distance between Gambles A and B, $\overline{AB}$, is computed as follows:

$$\overline{AB} = \sqrt{(\text{std.payoff}_B - \text{std.payoff}_A)^2 + (\text{std.prob}_B - \text{std.prob}_A)^2},$$

where std.payoff_A is the standardized payoff of Gamble A and std.prob_A is the standardized probability of winning of Gamble A. The computed distances represent how similar the given gambles are, which provides off-diagonal elements for the distance matrix D.

The last pieces of the specification of drift rate are priors for $\eta, \rho,$ and σ. These parameters account for the maximum covariance, the rate of decline in covariance with distance, and the extra variance, respectively. All parameters enter the model squared, so I define priors for squared parameters, instead of raw parameters as is. I use exponential priors for all three parameters as below,

$$\eta^2 \sim \text{Exp}(\lambda_\eta),$$
$$\rho^2 \sim \text{Exp}(\lambda_\rho),$$
$$\sigma^2 \sim \text{Exp}(\lambda_\sigma),$$

which completes the specification of the model for drift rate v, or subjective values $u_A, ..., u_E$.

Non-decision time t_{er} Non-decision time parameter t_{er} accounts for the time taken for all the processes that are not specified as the diffusion process, such as time for encoding information, or motor responses (e.g., pressing keyboard or mouse). This parameter is not of particular interest in this study, as no theoretical implications have been found for t_{er} in the field of economic decision theory. If anything, aging appears to have an effect on this parameter; a couple of studies about aging have found that older adults (age 60-75) in general showed slower t_{er}, compared to younger adults (college students), in lexical decision, or recognition memory tasks (Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Thapar, & McKoon, 2004). Although the current study doesn't aim to investigate any aging related effects, we still need to specify priors for t_{er} to initiate the Bayesian analysis of the diffusion model. Hence, I consider the following prior for t_{er}:

$$\begin{aligned} t_{er(p)} &= \Phi(\pi_{(p)}) \times \min(\mathrm{rt}_p), \quad p \in \{1, ..., P\}, \\ \pi_{(p)} &\sim \ \text{Normal}(\mu_{t_{er}}, \sigma^2_{t_{er}}), \\ \mu_{t_{er}} &\sim \ \text{Normal}(0, 1), \\ \sigma_{t_{er}} &\sim \ \text{Exp}(1) \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative normal distribution function and $\min(\mathrm{rt}_p)$ is the minimum RT of participant p. Since the range of $\Phi(\cdot)$ is (0, 1), the above prior ensures $t_{er(p)}$ to be smaller than $\min(\mathrm{rt}_p)$. Note that in this formulation, $\min(\mathrm{rt}_p)$ is used to specify the prior for $t_{er(p)}$. That is, the information obtained from data has been used to construct the prior, which goes against the philosophy of Bayesian analysis: priors represent prior beliefs about the parameters of interest before we see data. However, the current diffusion model is quite vulnerable to misspecification of priors. Specifically, it's non-decision time parameter that needs extra care when setting its prior, because even a slightly misspecified non-decision time prior may result in serious convergence issues, or crashes of the entire estimation process. A common problem that routinely occurs with this parameter is that its prior estimates non-decision time to be larger than its minimum observed RT, or a specific portion of RT that the non-decision time must not exceed. Thus, in practice, the minimum observed RT is often employed to specify the prior for t_{er} for reliable performance of the MCMC sampler (Donzallaz, Haaf, & Stevenson, 2022). Also the idea of using data for prior specification is not completely new, but has been around by the name of "Empirical Bayes" (Casella, 1985). Thus, I employ $\min(\mathrm{rt}_p)$ to construct the prior for $t_{er(p)}$, and the entire model specification is now completed.

Chapter 4

Simulation study: Parameter recovery simulations

In this chapter, I conduct a simulation study, where interest lies in recovering the data-generating parameters of the diffusion model. The diffusion model is one of the sequential sampling models that describe the decision-making process as an information accumulation process, also known as the diffusion process. Mathematically, the diffusion process involves stochastic differential equations, which often lead to intractable integrals, so simply estimating parameters of the diffusion model presents a enough challenge for applied researchers (Schwarz, 2001; Vandekerckhove & Tuerlinckx, 2007). Acknowledging such difficulties in the application of the diffusion model in practice, a number of software packages have been developed exclusively for diffusion model analysis (e.g., Vandekerckhove & Tuerlinckx, 2007; Voss & Voss, 2007; Wiecki et al., 2013), but now careless applications of the diffusion model, especially with regard to its across-trial variability in main parameters, are often considered to be a problem (see Boehm et al., 2018, for a discussion). Therefore, in this chapter, I examine the behavior of the particular version of the diffusion model I employ, from the aspect of parameter recovery.

4.1 Background

4.1.1 Diffusion model

The standard Ratcliff diffusion model (Ratcliff, 1978) has seven parameters, of which the drift rate v, decision criterion a (or boundary separation), starting point z, and non-decision time t_{er} are main parameters and the across-trial variability in drift rate s_v, starting point s_z and non-decision time $s_{t_{er}}$ are variability parameters that account for how much the main parameters vary from trial to trial. When the diffusion model is applied with all these seven parameters (which I refer to as the full diffusion model), the model becomes versatile enough to account for many benchmark phenomena in cognitive psychology (e.g., fast errors), but at the same time it becomes quite hard to estimate the parameters of the model reliably. Indeed, some simulation studies reported that the full diffusion model was not able to recover the across-trial variability in drift-rate and starting point specifically (Ratcliff & Tuerlinckx, 2002; van Ravenzwaaij & Oberauer, 2009). If the data are not enough, there is even a possibility that the full diffusion model can't recover the main parameters (van Ravenzwaaij, Donkin, & Vandekerckhove, 2017; Wagenmakers et al., 2007).

A straightforward way to address such estimation problems is to set the across-trial variability to 0. The motivation for this approach is that most studies' goals lie in the main parameters and these parameters can be estimated to a great precision even without including across-variability (Boehm et al., 2018; Lerche & Voss, 2016). One notable example is the EZ-diffusion model (Wagenmakers et al., 2007). The EZ-diffusion model sacrifices the versatility of the original diffusion model by forgoing across-trial variability (i.e., all across-trial variability is set to 0) and starting point (i.e., the starting point is set to the point equidistant from both decision criteria) for better performance on estimating the drift rate and decision criterion. Without the across-trial variability and starting point, the EZ-diffusion model is not able to provide a reliable prediction for RT distributions for errors, but the models' ability to estimate the drift rate and decision criterion has proven effective in many empirical settings (e.g., Enkavi et al., 2019; Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007). Plus, its closed-form equations allow the estimation of the parameters to be quick and easy, requiring no additional iterative fitting algorithms (Wagenmakers, van der Maas, et al., 2008). Recall, however, that the benefits of the EZ-diffusion model doesn't come free of charge. The EZ-diffusion model gains its robustness and tractability by sacrificing the starting point and across-trial variability in the parameters. If those forgone parameters are actually present in the data (i.e., across-trial variability of the parameters does exist, or the starting point *is* biased for some tasks in the given data), the EZ-diffusion model may produce biased estimates for main parameters, or as Ratcliff warned, they can even be wrong (Ratcliff, 2008).

The takeaway here is that it is crucial to understand the limit of the model in use. One cannot blindly let go of some parameters, or recklessly apply the full diffusion model without considering the setting where the model is being applied. Once the final model is settled, it is necessary to investigate what the model can do and cannot given the setting the model is to be applied. And a straightforward way to achieve this is to conduct simulation studies (Tuerlinckx, Maris, Ratcliff, & De Boeck, 2001). Which provides the primary motivation for the current simulation study, but before I proceed, I give a brief background of preferential choices in the next section, which serve as the main type of data for the current analysis.

4.1.2 Preferential choices

Although the diffusion model has been applied almost exclusively to perceptual tasks (e.g., moving dot task), the type of task of primary interest in this study is preferential choice tasks. Preferential choices are different from perceptual choices in a number of ways. The major difference is how the quality of a response is defined (Dutilh

& Rieskamp, 2016). In perceptual tasks, there is a clear decision criterion against which the given response is compared, producing a correct or wrong response. In a brightness discrimination task, for example, a participant is asked to determine which square between the two on screen is brighter as fast or/and accurate as possible. Any given answer is then compared against the criterion and recorded as correct or wrong. For perceptual tasks, such criterion is provided externally, mostly by the experimental design. In preferential choice tasks, however, there is no such criteria against which choices are compared. For example, if one prefers an apple to a banana in a food choice task, it wouldn't make sense to say the given preference is correct or wrong. All that matters is instead whether the decision maker is satisfied with his or her choices, or put differently, whether the choice is consistent with the decision maker's subjective value. Those criteria are usually generated internally, such as, from related memories or from personal values (Shepsle & Bonchek, 1997). Hence, various relations between accuracy and RT in cognitive psychology are no longer valid for preferential choices. The popular speed-accuracy tradeoff phenomenon makes no sense for preferential choice tasks, simply because there are no correct or wrong answers for preferential choices; if anything, the opposite relation has been reported, that is, fast responses likely lead to responses consistent with one's subjective values (Fudenberg et al., 2018).

For such reasons, many cognitive psychologists have considered preferential choices not suitable for the diffusion model analysis (Ratcliff & McKoon, 2008). But recent neurophysiological findings suggest that our brain would go through similar information accumulation processes for both perceptual *and* preferential choices (Basten et al., 2010; Krajbich et al., 2015; Milosavljevic et al., 2010; Polanía et al., 2014; Summerfield & Tsetsos, 2012). For example, Basten and her colleagues (2010) used fMRI and found the evidence that when in need of value-based decisions, such as preferential choices, our brain processes cost-related and benefit-related information in amygdala and ventral striatum, respectively, in the ventromedial prefrontal cortex. Then, the computed cost-benefit differences are found to accumulate toward a decision criterion in parietal cortex. In another study that examined EEG recorded during perceptual and value-based decision-making tasks, Polanía and his colleagues (2014) found that both types of choices went through the same kind of accumulation process that occurred in the same area of the brain.

In sum, an increasing number of neural evidence indicates that the same brain mechanisms that underlie perceptual choices are also used for value-based decisions. For both types of tasks, a group of related neurons are found to gradually increase their firing rates, suggesting the accumulation of information, and a decision is made when the firing rate of those neurons exceeds a certain criterion (Bogacz et al., 2010; Gold & Shadlen, 2007; Schall, 2001). Such findings have greatly motivated the application of the diffusion model analysis to the preferential choice data, which in turn has inspired the current simulation study. Here is how this chapter is organized. In what follows, I first walk through the exact steps I took to generate data, which includes what software package I use and the actual parameter values that are used to generate the data. Then, I specify the model that I fit against the simulated data to recover the data-generating parameters. The model I use here is mostly same as the one I introduced in Chapter 3, so I encourage to consult Chapter 3 for a detailed account of the model. Finally, I conclude the chapter with the results.

4.2 Simulating data

In this simulation study, I simulate two sets of data from two different diffusion models: the full diffusion model (with seven parameters) and the diffusion model with the constraints specified in Chapter 3, which I refer to as the constrained diffusion model. This is due mainly to the argument made by Ratcliff (2008) that when across-trial variability is present in the data, diffusion models without across-trial variability (e.g., EZ-diffusion model) can produce wrong estimates for the main parameters. But, the opposite finding has also been reported; Lerche and Voss (2016) conducted simulation studies, where they found that the diffusion model without across-trial variability was well able to recover the data-generating parameters of the full diffusion model. Those mixed results indicate that the benefit of including the across-trial variability in the diffusion model analysis is still debatable at least for the estimation of the main parameters. Hence, in this simulation, I generate two data sets, one from the full diffusion model with all seven parameters included and one from the constrained diffusion model, where the across-trial variability and starting point parameters are not included. Then, I fit the constrained diffusion model to both simulated data sets to see if the given model is able to recover the true data-generating parameters of both models.

The specific design of the current simulation study generally follows Boehm et al.'s (2018) procedure. Specifically, I follow the procedure of the level-3 simulation study, where the data-generating parameters form hierarchical structures, requiring one to use a hierarchical model to recover group-level parameters. This data set will obviously put the performance of the hierarchical structure of the current model to the test.

4.2.1 Simulating data from the full diffusion model

Before I proceed with simulating data sets, I first discuss a few details of the current simulation study. First, I employ the paradigm of Tversky's (1969) experiments, where five monetary gambles are used to generate binary choice data; that is, 10 unique gamble pairs are used (i.e., $\binom{5}{2} = 10$). The gambles differ in payoff and probability of winning in a way that the two attributes tradeoff with each other (i.e., the more the payoff the less the probability of winning; see Figure 4.1). The

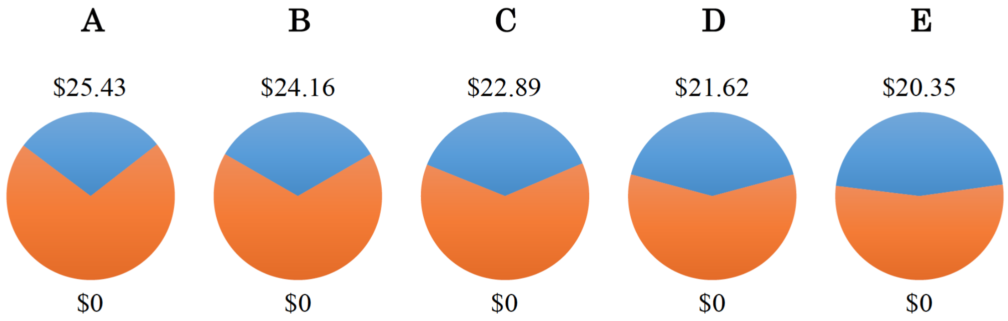


Figure 4.1: Gamble set used to simulate the data. Gambles are in the pie format, where the blue-colored area represents the probability of winning the specified payoff. Different gambles have different values of payoff and probability of winning, in a way that the two attributes tradeoff with each other.

same gamble pairs appear multiple times in an experimental session to grasp inherent variability of preferential choices (Tversky, 1969). In this simulation, I assume 12 repetitions for each of the gamble pairs, which produce $10 \times 12 = 120$ trials per individual. Second, drift rates are determined by subjective value difference between the two monetary gambles. Since there are 10 different gamble pairs, 10 different drift rates are supposed to be estimated for each participant. However, due to the way the drift rates are determined (specified in Chapter 3), estimating just 5 subjective values, each value for each gamble, will suffice. Finally, the current simulation study has the standard deviation of the diffusion process, or the scale parameter, fixed at 1. The typical value in the literature is 0.1 (Ratcliff, 1978; Vandekerckhove & Tuerlinckx, 2007), but for mathematical convenience, some diffusion models have deliberately fixed its scale parameter at 1 (Navarro & Fuss, 2009; Voss et al., 2004). The current simulation follows the latter case, and so, estimates of v, a, and z are 10 times larger in size compared to the corresponding estimates for Ratcliff's diffusion model.

The full diffusion model has seven parameters: drift rate v, decision criterion a, starting point z, non-decision time t_{er}, across-trial variability in drift rate s_v, across-trial variability in starting point s_z, across-trial variability in non-decision time $s_{t_{er}}$.

As in Boehm et al.'s (2018) level-3 simulation study, only the main parameters are assumed to be drawn from the corresponding group-level distributions, while across-trial variability parameters remain the same across participants. As discussed above, the drift rate parameter v is not directly estimated from the data. Instead, subjective value of each gamble $u_A,\ u_B,\ u_C,\ u_D,\ u_E$ is estimated first, then drift rates are computed using the estimated subjective values.

The specific values of data-generating parameters are listed in Tables 4.1 and 4.2. In this simulation study, I consider four classifications, resulting from combining decision-making styles (conservative or less-conservative) and risk attitudes (risk-seeking or risk-averse). The reason I consider such classifications is that each of the classifications supposedly gives rise to a different data pattern and I would like to put the model to the test against various patterns of data. Specifically, conservative participants have higher decision criterion than less-conservative ones; risk-seeking participants prefer gambles with high payoff, while risk-averse participants prefer gambles with high probability of winning. Each classification provides a unique profile of the parameters, where decision-making styles affect the decision criterion parameter a and risk attitudes affect the way subjective values of the gambles $u_A, ..., u_E$ are determined. Table 4.1 shows group-level parameters for each of the classifications.

Table 4.1: Group-level parameters

	Classification		Group-level parameters k							
	Risk attitudes	Decision criterion	a	u_A	u_B	u_C	u_D	u_E	t_{er}	z
Mean μ_k	Risk-seeking	Conservative	2	5	4	3	2	1	0.43	0.5
	Risk-seeking	Less-Conservative	1	5	4	3	2	1	0.43	0.5
	Risk-averse	Conservative	2	1	2	3	4	5	0.43	0.5
	Risk-averse	Less-Conservative	1	1	2	3	4	5	0.43	0.5
Standard deviation σ_k			0.5	0.7	0.7	0.7	0.7	0.7	0.1	0.04

Note. Group-level parameters are used to generate individual-level parameters for each classification. Individual-level parameters are drawn from the corresponding normal distributions with mean of μ_k and standard deviation of σ_k, where the index k runs through the listed parameters. Standard deviations remain the same across classifications.

Table 4.2: Individual-level parameters

	Classification		Individual-level parameters							
	Risk attitudes	Decision criterion	a	u_A	u_B	u_C	u_D	u_E	t_{er}	z
PP1	Risk-seeking	Conservative	1.72	4.61	4.32	3.28	2.49	1.39	0.30	0.52
PP2			1.88	4.84	3.11	3.08	1.67	0.96	0.65	0.48
PP3			2.78	6.09	3.52	2.61	1.25	0.79	0.55	0.49
PP4			2.04	5.05	3.69	4.25	1.85	0.73	0.32	0.46
PP5		Less-conservative	1.09	4.85	4.70	4.53	1.83	1.36	0.44	0.51
PP6			0.93	5.23	4.38	4.07	1.76	1.21	0.34	0.50
PP7			1.00	5.77	4.17	2.84	1.33	1.07	0.38	0.46
PP8			1.19	5.30	3.56	2.28	1.97	0.55	0.40	0.50
PP9	Risk-averse	Conservative	2.39	0.93	2.59	3.27	5.40	3.59	0.60	0.54
PP10			2.38	0.05	2.67	2.81	4.05	5.15	0.52	0.45
PP11			2.17	0.53	2.48	3.08	5.31	5.87	0.45	0.47
PP12			1.50	1.34	1.02	3.09	3.05	6.43	0.55	0.56
PP13		Less-conservative	1.55	0.76	2.12	2.85	3.50	3.79	0.38	0.45
PP14			1.04	1.06	2.05	3.12	3.47	4.58	0.50	0.49
PP15			1.38	2.12	2.30	3.82	3.34	4.75	0.42	0.53
PP16			0.75	0.94	2.02	3.74	3.26	5.49	0.49	0.50
Across-trial variability parameters:			$s_v = 1.6,\ s_{t_{er}} = 0.15,\ s_z = 0.3$							

Note. Each value shown in this table enters the diffusion model to generate the data. To generate the data from the full diffusion model, the main parameters are allowed to vary from trial to trial, according to the across-trial variability parameters. To generate the data from the constrained diffusion model, the across-trial variability is all set to 0 and the starting point is fixed at 0.5 for all participants. For the rest of the parameters, the same values are used for both models to generate data.

The group-level parameters serve as the mean and standard deviation of normal distribution, respectively, which generates values of individual-level parameters for each classification (the generated values are listed in Table 4.2). In the current simulation, the three across-trial variability parameters s_v, $s_{t_{er}}$, and s_z remain the same across participants.

To simulate data from the full diffusion model (to which I refer as the full data), I use the `RWiener` (Wabersich & Vandekerckhove, 2014) `R` package (R Core Team, 2022). This package generates data using the Wiener first-passage time distribution (Navarro & Fuss, 2009). Note that the `RWiener` package is only concerned with the Wiener process (i.e., the process where information accumulates over time until it reaches one of the criteria), not with parametric assumptions, such as across-trial variability. In order to incorporate across-trial variability in generating data, package users have to provide its main function `rwiener` with different values of parameters each trial. Hence, as in the standard Ratcliff diffusion model, I use uniform distributions to draw values for starting point and non-decision time every trial. Specifically, the mean of each uniform distribution is set to z and t_{er}, respectively, with its lower bound and upper bound set to $[z - s_z/2, z + s_z/2]$ and $[t_{er} - s_{t_{er}}/2, t_{er} + s_{t_{er}}/2]$, so that its range equals s_z and $s_{t_{er}}$, respectively.

For the drift rate, I use normal distribution with mean of v and standard deviation of s_v. Recall that the mean drift rates are determined by the difference in subjective value between the two given alternatives. For example, if the gamble pair A and B are present on trial j, mean drift rate v_j is given by $u_A - u_B$. So, what matters is the difference between any two of subjective values, not the individual value of itself. Hence, when generating data, I first center the subjective values by subtracting the subjective value of Gamble C from each subjective value, and use the centered values to generate data. For example, Participant 1's subjective values are 4.61, 4.32, 3.28, 2.49, 1.39 for Gambles A, B, C, D, E as listed in Table 4.2. After the centering, those values become 1.33, 1.04, 0, -0.79, -1.89, and drift rates are computed using these centered values. Since the value of Gamble C is fixed at 0 for all participants, we only need to consider values of Gambles A, B, D, E, which effectively solves the identifiability problem of the model.

4.2.2 Simulating data from the constrained diffusion model

The second data set is simulated from the constrained diffusion model (to which I refer as the constrained data), where the across-trial variability and starting point are fixed. This is the model I use to recover the data-generating parameters of the full and constrained diffusion models, so the current data set provides a direct test of the model's ability to recover its parameters. In this model, all across-trial variability is set to 0 and starting point is fixed at 0.5 (i.e., unbiased choices) for all participants. The hierarchical structures are still assumed for the rest of the main parameters, i.e., decision criterion, and drift rate, and non-decision time, so the same values in Table 4.2 are used to generate data, except that the across-trial variability and starting point are now fixed at 0 and 0.5, respectively.

Again, I use the `RWiener` (Wabersich & Vandekerckhove, 2014) `R` package to simulate the data from the constrained diffusion model. Since there is no across-trial variability assumed for any parameters, the values listed in Table 4.2 directly enter the main data-generating function `rwiener`. As before, the subjective value of Gamble C is subtracted from each subjective value, and the centered subjective values are used to generate data. The primary interest of the current simulation study centers on the parameters of decision criterion a and subjective values of gambles $u_A, ..., u_E$. These parameters are particularly related to the main hypotheses of the present dissertation, hence the model's ability to recover those parameters is critical for the current study.

4.3 Model specification

As described in Chapter 3, the diffusion model I choose assumes no across-trial variability in its main parameters and has its starting point fixed at 0.5. Let $\boldsymbol{y}_{(pj)}$ be the response vector of person p on trial j, with its components being response $c_{(pj)}$ and response time $t_{(pj)}$. Then, $\boldsymbol{y}_{(pj)}$ are distributed as the Wiener first-passage time distribution (Navarro & Fuss, 2009) (abbreviated to Wiener) as follows:

$$\begin{pmatrix} c_{(pj)} \\ t_{(pj)} \end{pmatrix} = \boldsymbol{y}_{(pj)} \sim \text{Wiener}(a_{(p)}, z_{(pj)} = 0.5, t_{er(p)}, v_{(pj)}),$$

where p indexes people, $p = 1, ..., P$, j indexes trials, $j = 1, ..., J$. The four main parameters are denoted by a, z, t_{er}, and v, where a is the decision criterion, z is the starting point, t_{er} is the non-decision time, and v is the drift rate. Since we are using Bayesian methods to fit the model to the data, all the main parameters, except for z, require priors to be specified. Here, the priors largely follow the setup I described in Chapter 3, so for those who are curious about reasons or justifications for the choices I made for the priors, I encourage to consult Chapter 3.

The decision criterion a governs the amount of information needed for a decision. Since a must be positive in this model, I take a log-transformation of a and define the rest of the model on a log scale. This parameter is directly related to one's decision making style: conservative vs. less-conservative. Higher values of a are associated with conservative decision-making style, while lower values of a are associated with less-conservative decision-making style. Every participant is assumed to have a unique value of this parameter, and since no across-trial variability is assumed for a, the same value of a carries on throughout the whole experimental session. In this simulation study, we know which participant belongs to which group prior to the analysis, so this information enters the model as an explanatory variable for a. Let $X_{a(p)}$ be the group membership, where $X_{a(p)} = 1$ if participant p is conservative, $X_{a(p)} = 0$ otherwise. Then, the prior for a is given by:

$$log(a_{(p)}) \sim \text{Normal}(\mu_{a(p)},\ \sigma_a^2),\ \ p \in \{1, ..., P\},$$

$$\text{where } \mu_{a(p)} = \beta_{0a} + \beta_{1a} X_{a(p)}.$$

And its hyperpriors are given by:

$$\beta_{0a} \sim \text{Normal}(0.5,\ 0.5^2),$$
$$\beta_{1a} \sim \text{Normal}(0,\ 0.5^2),$$
$$\sigma_a \sim \text{Exp}(0.5).$$

For the priors of β_{0a} and β_{1a}, I choose normal distributions. In many applications of the diffusion model, decision criterion a is usually estimated to be around 1 (Ratcliff, 2002; Ratcliff, Gomez, & McKoon, 2004; Wagenmakers, Ratcliff, et al., 2008),[1] and to find out what specific values of mean and standard deviation for β_{0a} and β_{1a} would give rise to a of 1, I examined prior predictive distribution of a. This is a great way to decide priors especially when parameters are transformed before being modeled. As a result, I decide to go with mean of 0.5 and standard deviation of 0.5 for β_{0a} and mean of 0 and standard deviation of 0.5 for β_{1a}, with σ_a being distributed as Exp(0.5). The prior predictive distribution of a with these values is shown in Figure 4.2. As we can see, the majority of probability density of prior predictive distribution of a is centered around 1, consistent with many empirical estimates of a. Since I suppose no group difference *a priori*, the two distributions in Figure 4.2 are substantially overlapped with each other. If the data suggest any difference between the two groups, one of the posterior densities of a will be shifted away from its prior mean toward either a positive or negative direction, as the data suggest.

Drift rate v determines the mean rate of information uptake of the diffusion process. In other words, v governs how fast the accumulator accumulates information toward one of the decision criteria. Recall that drift rates are given by subjective value difference between any pair of gambles. Thus, it's subjective values that need

[1] The current diffusion model has its scale parameter s of the diffusion process set to 1. The standard Ratcliff diffusion model's scale parameter is 0.1, so the current diffusion model's estimates are 10 times larger in size compared to the corresponding parameters of the standard Ratcliff diffusion model.

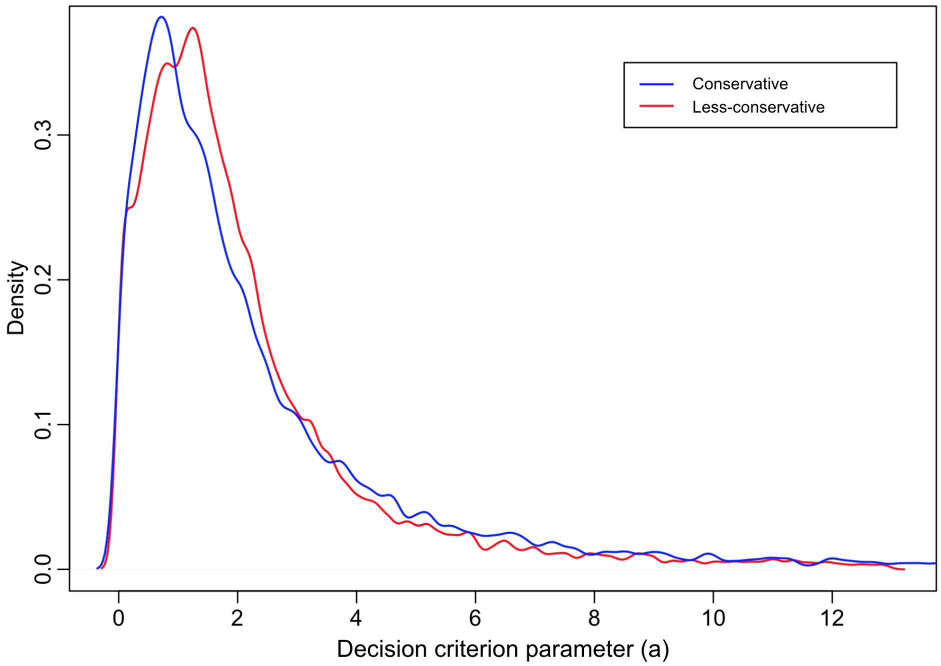


Figure 4.2: Prior predictive distributions of a for conservative and less-conservative groups. Majority of probability densities of both distributions are centered around 1.

priors, not the drift rate *per se*. Let $v_{AB(p)}$ be the drift rate of participant p when the gamble pair A and B is present. Also, let $u_{A(p)}$, ..., $u_{E(p)}$ be the subjective values of gambles A through E of participant p. Then, drift rate $v_{AB(p)}$ for the gamble pair A and B is given by:

$$v_{AB(p)} = u_{A(p)} - u_{B(p)}.$$

Note that for identifiability of the model, we fix the subjective value of Gamble C at 0 and estimate the rest of the subjective values relative to Gamble C. Let $\boldsymbol{u}_{(p)}$ be a vector of subjective values of participant p, i.e., $\boldsymbol{u}_{(p)} = (u_{A(p)}, u_{B(p)}, u_{D(p)}, u_{E(p)})^T$. Then, the prior for $\boldsymbol{u}_{(p)}$ is set as follows:

$$\begin{pmatrix} u_{A(p)} \\ u_{B(p)} \\ u_{D(p)} \\ u_{E(p)} \end{pmatrix} = \boldsymbol{u}_{(p)} \sim \text{MVN}_4 \left(\begin{pmatrix} \mu_{A(p)} \\ \mu_{B(p)} \\ \mu_{D(p)} \\ \mu_{E(p)} \end{pmatrix}, \ \Sigma \right).$$

The above prior has its own parameters, i.e., $(\mu_A, \mu_B, \mu_D, \mu_E)^T$ and Σ, that require priors to be specified. First, we begin by setting priors for the mean vector of subjective values, and in doing that, we make use of the risk attitude information of each participant. Recall that the current simulation study incorporates participants' risk attitudes in generating data. Risk-seeking participants value gambles with higher payoff more than gambles with higher probability of winning; risk-averse participants value gambles with higher probability of winning more than gambles with higher payoff. Since we know which participant is risk-seeking, and which participant is risk averse, prior to the analysis, we should let its priors reflect such different risk attitudes across participants. Let $X_{u(p)}$ be the classification of risk attitude of participant p, where $X_{u(p)} = 1$ if participant p is risk-seeking, $X_{u(p)} = 0$ otherwise. Then, $X_{u(p)}$ enters the model as an explanatory variable as follows:

$$\mu_{k(p)} = \beta_{0k} + \beta_{1k} X_{u(p)}, \quad k \in \{A, B, D, E\}.$$

For β_{0k} and β_{1k}, I employ weakly informative hyperpriors. Since subjective values of the gambles are estimated relative to gamble C, the hyperpriors should be centered around 0:

$$\beta_{0k} \sim \text{Normal}(0, 3^2), \ \beta_{1k} \sim \text{Normal}(0, 3^2), \quad k \in \{A, B, D, E\}.$$

I again choose normal distributions for the hyperpriors, because normal distributions are the ones that maximize entropy when no information, other than its variance, is available (see Chapter 3 for more accounts of this point). Their means are set to 0 and variance is 3, reflecting that drift rates are usually estimated to be from -2 to 2 in typical empirical settings (Matzke & Wagenmakers, 2009).

When it comes to the prior for Σ, I employ the prior for variance-covariance matrix by McElreath (2020), where he used geographic distance between two objects to compute their covariance. Let Σ_{ij} be the covariance between any pair of gambles i and j, which is given by:

$$\Sigma_{ij} = \eta^2 \text{exp}(-\rho^2 D_{ij}^2) + \delta_{ij}\sigma^2, \ \ i, j \in \{A, B, D, E\},$$
$$\text{where } \delta_{ij} = \begin{cases} 1 \ \text{ if } i = j, \\ 0 \ \text{ if } i \neq j. \end{cases}$$

The above formula has three parameters η, ρ, and σ that need priors to be specified. The three parameters account for the maximum covariance, the rate of decline in covariance with distance, and the extra variance, respectively. All parameters enter the model squared, so I define priors for squared parameters. I use exponential priors for all three parameters as below:

$$\eta^2 \sim \text{Exp}(2),$$
$$\rho^2 \sim \text{Exp}(1),$$
$$\sigma^2 \sim \text{Exp}(0.5).$$

Non-decision time t_{er} accounts for the time taken for all the processes that are not the diffusion process. One notable feature of the prior for t_{er} is the inclusion of the observed minimum RT. Theoretically, priors formulate prior beliefs about the parameters of interest before the data are given, so we shouldn't use any feature of observed data to construct priors. But, the current model includes the observed minimum RT in the prior for t_{er} for the following reasons. First, the diffusion model requires that t_{er} be smaller than the observed minimum RT, which is, without the use of the minimum RT, quite hard to meet. And second, because of the requirement, the sampling process is vulnerable to prior values of t_{er}, which often leads to a collapse of the entire Markov chains if we don't limit the prior values smaller than the minimum RT. Thus, the observed minimum RT is used for the prior of t_{er} primarily for a reliable performance of the MCMC sampler.

Each participant is assumed to have a unique value of t_{er}, with no across-trial variability assumed. So, t_{er} remains the same throughout the entire experiment. A hierarchical structure is imposed on the prior of t_{er} to produce better estimates for each participant. I consider the following prior for t_{er}: for $p \in \{1, ..., P\}$,

$$\begin{aligned} t_{er(p)} &= \Phi(\pi_{(p)}) \times \min(\mathrm{rt}_p), \\ \pi_{(p)} &\sim \ \mathrm{Normal}(\mu_{t_{er}}, \sigma^2_{t_{er}}), \end{aligned}$$

and its hyperpriors are:

$$\begin{aligned} \mu_{t_{er}} &\sim \ \mathrm{Normal}(0, 1), \\ \sigma_{t_{er}} &\sim \ \mathrm{Exp}(1), \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative normal distribution function and $\min(\mathrm{rt}_p)$ is the minimum RT of participant p. Since the range of $\Phi(\cdot)$ is bounded between 0 and 1, the above prior ensures that $t_{er(p)}$ is smaller than $\min(\mathrm{rt}_p)$.

4.4 Results

The specified model is fitted to both of the two simulated data sets, one from the full diffusion model and one from the constrained model, using Stan, a platform for general statistical modeling and computation (Stan Development Team, 2022). Stan is arguably the most powerful program for full Bayesian analysis, with its state-of-art sampling algorithm. Particularly, Stan implements the Hamiltonian Monte Carlo (Duane, Kennedy, Pendleton, & Roweth, 1987; Neal, 2011), and its extension, the No-U-Turn (NUTS) Sampler (Hoffman & Gelman, 2014). The sampling algorithm in Stan provides much more efficient performance, compared with other conventional MCMC samplers, such as Metropolis-Hasting algorithm, or Gibbs sampler, in that Stan's algorithm converges much faster toward the target density (Hoffman & Gelman, 2014). For models with highly correlated parameters, which include the diffusion model, Stan easily outperforms the Gibbs sampler as Stan produces less auto-correlated samples (Annis, Miller, & Palmeri, 2017). In addition, Stan includes the Wiener first-passage time distribution (Navarro & Fuss, 2009) as one of its default probability density functions, which makes the diffusion model analysis easier than it has ever been.

In this simulation study, I follow the default settings recommended by Stan to generate the posterior samples. In particular, for both data sets, four chains are used to generate posterior samples; from each chain, I generate 1,000 posterior samples, after 1,000 samples of warmup. During the sampling process, a single divergent transition (Betancourt, 2016) was not observed in any chains, and after the sampling is done, convergence of the chains was confirmed by trace plots, effective sample sizes, and Gelman and Rubin diagnostics (Gelman & Rubin, 1992). Thus, I conclude that the resulting posterior samples well represent the target posterior density, and so I make all the statistical inferences that I need for the current simulation study based on these 4,000 posterior samples.

4.4.1 Result: The constrained data

I first discuss the results of the analysis of the data generated from the constrained diffusion model. Since this data set is generated from the same model I aim to fit to data, it allows us to test the given model's ability to recover its parameters in a direct way. Hence, I first examine the model's performance in the setting where it should be performing well. Then, I generalize the setting, where I assume the values of the main parameters vary from trial to trial and the starting point is no longer fixed at 0.5, and see how well the given model is capable of recovering all the main parameters.

I begin by evaluating the model fit against the data. One way to achieve this is to use the posterior predictive density (Myung et al., 2005). The posterior predictive density describes the density of the *future* data $\boldsymbol{y}^{\text{pred}}$ predicted by the given model, if we hypothetically replicate the same experiment that produces the current data $\boldsymbol{y}$. So, we can measure how well the given model describe the observed data, by comparing the predicted future data against the observed data. If the model provides an adequate fit to the data, its future data should be well consistent with the observed data. The posterior predictive density can be obtained via integrating out all parameters $\boldsymbol{\theta}$ of the given model. Specifically, the posterior predictive density $p(\boldsymbol{y}^{\text{pred}}|\boldsymbol{y})$ is given by:

$$p(\boldsymbol{y}^{\text{pred}}|\boldsymbol{y}) = \int_{\Theta} p(\boldsymbol{y}^{\text{pred}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}, \tag{4.1}$$

where Θ is the parameter space.

In practice, it might be impossible to evaluate the above integration for most of the problems. But, thanks to the MCMC sampling algorithms, we are now able to sample values of the parameters from its posterior densities, and with those posterior samples, the posterior predictive density can be easily obtained using the Monte Carlo methods. That is, we simply feed posterior samples of each parameter, sampled from its marginal posterior density, to the data generating process to predict the future data. Below, I list the steps I take to generate the posterior predictive samples:

1. Generate posterior samples of the parameters from its respective marginal posterior density: for current model, the parameters are a, $\boldsymbol{u} = (u_A, ..., u_E)$, and t_{er}, and so its marginal posterior densities are $p(a|\boldsymbol{y})$, $p(\boldsymbol{u}|\boldsymbol{y})$, and $p(t_{er}|\boldsymbol{y})$, respectively

2. Randomly choose a number of generated posterior samples for each parameter

3. Feed the chosen values to the data-generating process to generate the data

The MCMC technique we employ to estimate the model takes care of the first step. As mentioned earlier, I use Stan to generate posterior samples of the diffusion model parameters; the number of the generated posterior samples is 1,000 per chain, resulting in 4,000 posterior samples in total. Then, I randomly choose 400 samples out of 4,000 and use the chosen values to generate the future data. The particular function I use to generate the data is `rwiener` from the `RWiener` (Wabersich & Vandekerckhove, 2014) `R` package, the same function I use to simulate the data. Note, however, that I only use 400 samples, instead of 4,000, to generate the data. This is simply because generating data from all of the 4,000 posterior samples requires much more of computer powers than I can afford, so I decide to go with a smaller number of posterior samples. I believe the posterior predictive density based on 400 generated data can still provide an adequate approximation of the posterior predictive density.

Since the diffusion model analyzes choices and response time simultaneously, the model fit is evaluated with respect to these two aspects of data. For choice data, I employ the Bayesian p-value (Myung et al., 2005), and for response time data, I compare the shape of the distribution of observed response time against predicted response time to see if we observe any significant difference between the two shapes of the distributions.

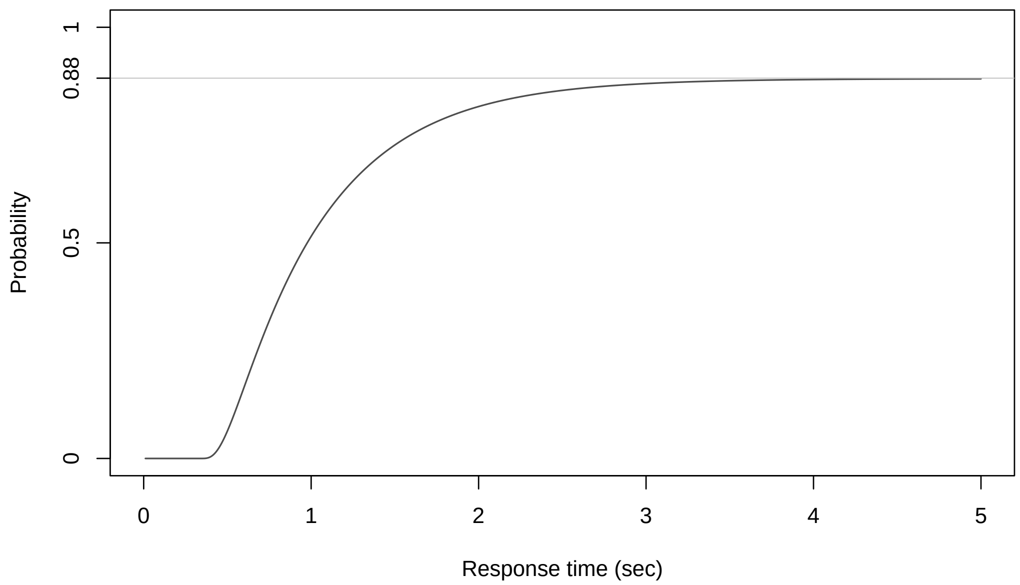


Figure 4.3: Cumulative density function for the diffusion model. To generate this CDF, the main parameters of the diffusion model are set as follows: $a = 2$, $t_{er} = 0.3$, $z = 0.5$, and $v = 1$. The CDF approaches toward its limiting value of 0.88 as response time increases.

First, in order to obtain the Bayesian p-value, I follow Myung et al. (2005) and use the generalized Pearson chi-square discrepancy function shown below:

$$\chi^2(\boldsymbol{y};\boldsymbol{p}) = \sum_{i=1}^{n} \frac{(y_i - N_i p_i)^2}{N_i p_i}, \tag{4.2}$$

where index i runs through the possible gamble pairs of five gambles, that is, $\{AB, AC, ..., DE\}$, so $n = 10$ in this case; y_i is the number of observed choices made for the left gamble of the given pair i (i.e., it refers to A when i equals AB); N_i is the number of repetition of gamble pair i; and p_i is the estimated probability of choosing the left gamble when gamble pair i is given.

Myung et al. (2005) proposed the use of the Pearson chi-square discrepancy function to test decision axioms, but the model of interest in the current study is the diffusion model. The major difference between the two models is that decision axioms directly model the binary choice probabilities, so it is easy to obtain posterior samples of p_i during its MCMC sampling, whereas the diffusion model doesn't include binary choice probabilities as its parameters in the model, so the MCMC algorithm of the diffusion model doesn't sample choice probabilities p_i during its sampling process. It only manifests itself via the main parameters of the diffusion model. So, I use the `pwiener` function from the `RWiener` package, which returns the probability of the model predicting the "upper" response (or equivalently, it refers to the left gamble between a given gamble pair for the current study) at a specified response time by computing the Cumulative Density Function (CDF) for the diffusion model. The `pwiener` function yields a different probability, if a different set of values of the main parameters and a different response time are given to the function. Note, however, that as response time increases, the probability of the "upper" response approaches toward its limiting value. For example, the diffusion model with decision criterion a of 2, non-decision time t_{er} of 0.3, starting point z of 0.5, and drift rate v of 1 produces the CDF shown in Figure 4.3, with its limiting probability of 0.88 as response time increases. So, I compute the limiting probability for each of the sampled parameter values, $\boldsymbol{\theta}^{(t)}_{t=1,\ldots,T}$, and use it as p_i.

Once we have Equation (4.2) computed using the limiting probability for the given values of the main parameters, the rest of the steps for the Bayesian p-value are straightforward to follow. Specifically, the Bayesian p-value is given by:

$$\text{Bayesian } p\text{-value} \equiv \Pr\{\chi^2(\boldsymbol{y}^{\text{pred}};\boldsymbol{p}) \geq \chi^2(\boldsymbol{y};\boldsymbol{p})\},$$

which can be approximated by applying the Monte Carlo method as follows:

$$\frac{1}{T}\sum_{t=1}^{T} I(\chi^2(\boldsymbol{y}^{\text{pred}(t)};\boldsymbol{p}^{(t)}) \geq \chi^2(\boldsymbol{y};\boldsymbol{p}^{(t)})),$$

where $I(\cdot)$ is a function that returns 1 if its argument is true, 0 otherwise. In this

Table 4.3: Bayesian p-values for simulation study

	Bayesian p-value			Bayesian p-value	
Participant	Constrained data	Full data	Participant	Constrained data	Full data
PP1	0.60	0.54	PP9	0.62	0.46
PP2	0.51	0.74	PP10	0.54	0.50
PP3	0.52	0.22	PP11	0.80	0.76
PP4	0.39	0.55	PP12	0.69	0.44
PP5	0.41	0.63	PP13	0.36	0.76
PP6	0.17	0.42	PP14	0.07	0.64
PP7	0.46	0.34	PP15	0.54	0.70
PP8	0.45	0.28	PP16	0.21	0.23

Note. Constrained data refers to the data set generated from the constrained diffusion model, and full data refers to the data set generated from the full diffusion model. None of the Bayesian p-values are smaller than 0.05, indicating that the constrained diffusion model provides an adequate fit to both data sets.

study, I obtain 400 posterior predictive samples for each of the main parameters of the diffusion model, so the Bayesian p-value is computed via the 400 Monte Carlo samples (i.e., $T = 400$). The results are shown in Table 4.3.

As with the frequentist p-value, a small Bayesian p-value, usually less than 0.05, indicates a lack of fit of the model against the data, yet a higher value doesn't necessarily imply a better fit of the model. I compute the Bayesian p-value for each of the participants, because different participants are fitted with diffusion models with different parameter values, so I seek to evaluate the diffusion model fit for all participants. As seen in Table 4.3, none of the computed Bayesian p-values is smaller than 0.05, indicating that the given diffusion model provides an adequate fit to both of the two simulated data sets. However, the Bayesian p-value of Participant 14 from the constrained data set is estimated to be 0.07, a relatively small value. After investigating the data and the prediction made by posterior predictive samples side by side, I conclude that the discrepancy between the observed and predicted choice proportions for Participant 14 doesn't provide evidence for a significant lack of fit of the model (see Table 4.4).

To evaluate the model fit in terms of response time (RT), I compare the density plot of the posterior predictive RT data against the observed RT. First, I collapse the observed RT data across gamble pairs and across participants to generate one plot of the whole RT data. Then, I superimpose the density plot of the predicted RT data, also collapsed across gamble pairs and participants, on top of the density plot of the observed data. To better distinguish one from another, the observed RT is displayed via a histogram, while the predicted RT is displayed via a density curve. Figure 4.4 shows the density of RT for the full and constrained data.

Figure 4.4 shows no visible gap between the observed and predicted RT. The given model accounts for the RT of the constrained data well, given that the density curve, generated by the posterior predictive samples, is overlapped with the histogram. To take a look further into RT, I create the same plot of a histogram with a density curve superimposed on it for each participant. This is because Participant 14 has a small value of the Bayesian p-value when we do the posterior predictive check for choice data, so I am interested in looking into each participant's RT. Figure 4.5 shows the observed RT as a histogram and the predicted RT as a density curve for each participant.

In Figure 4.5, there seem no significantly visible gaps between the observed and predicted densities across participants. Despite its small Bayesian p-value, Participant 14's RT is almost perfectly accounted for by the given model; other participants' RT are also well described by the given model. If anything, the model tends to predict the peak of RT density slightly faster than the observed RT (see Participants 2, 3, 9, and 10), but overall the model provides a good fit for the data.

Table 4.4: Observed and predicted choice proportions for Participant 14

	Choice proportions of gamble pairs									
	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
Observed proportions	0.17	0.00	0.17	0.08	0.33	0.25	0.00	0.33	0.25	0.00
Predicted proportions	0.18	0.07	0.08	0.01	0.26	0.24	0.01	0.50	0.09	0.08

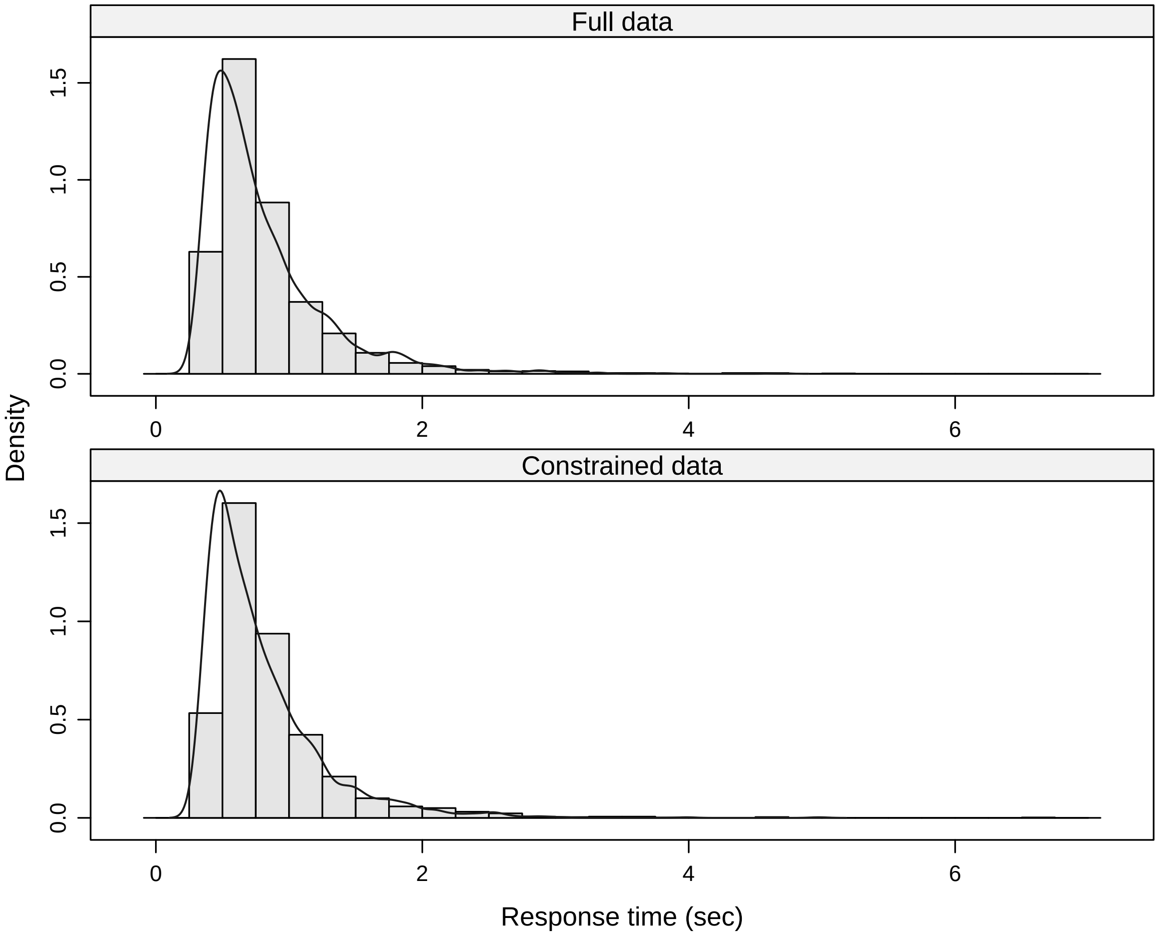


Figure 4.4: Density plots of response time for the full and constrained data. The RT data are collapsed across participants. The density of the observed RT is displayed via a histogram, and the density of the predicted RT is displayed via a density curve.

Now we turn our attention to the parameter recovery results. The main focus of this analysis is on the two main parameters: decision criterion a, and drift rates, or subjective values $u_A, ..., u_E$. Since the current data set has been generated using specific values of the parameters, the aim of the analysis is to recover its true data-generating parameters from the data. I use the posterior mean as a point estimate for each parameter, and its 95% credible intervals as a measure of the uncertainty as to the estimation. I first examine the group-level parameters.

Recall that the participants are classified into one of four classifications according

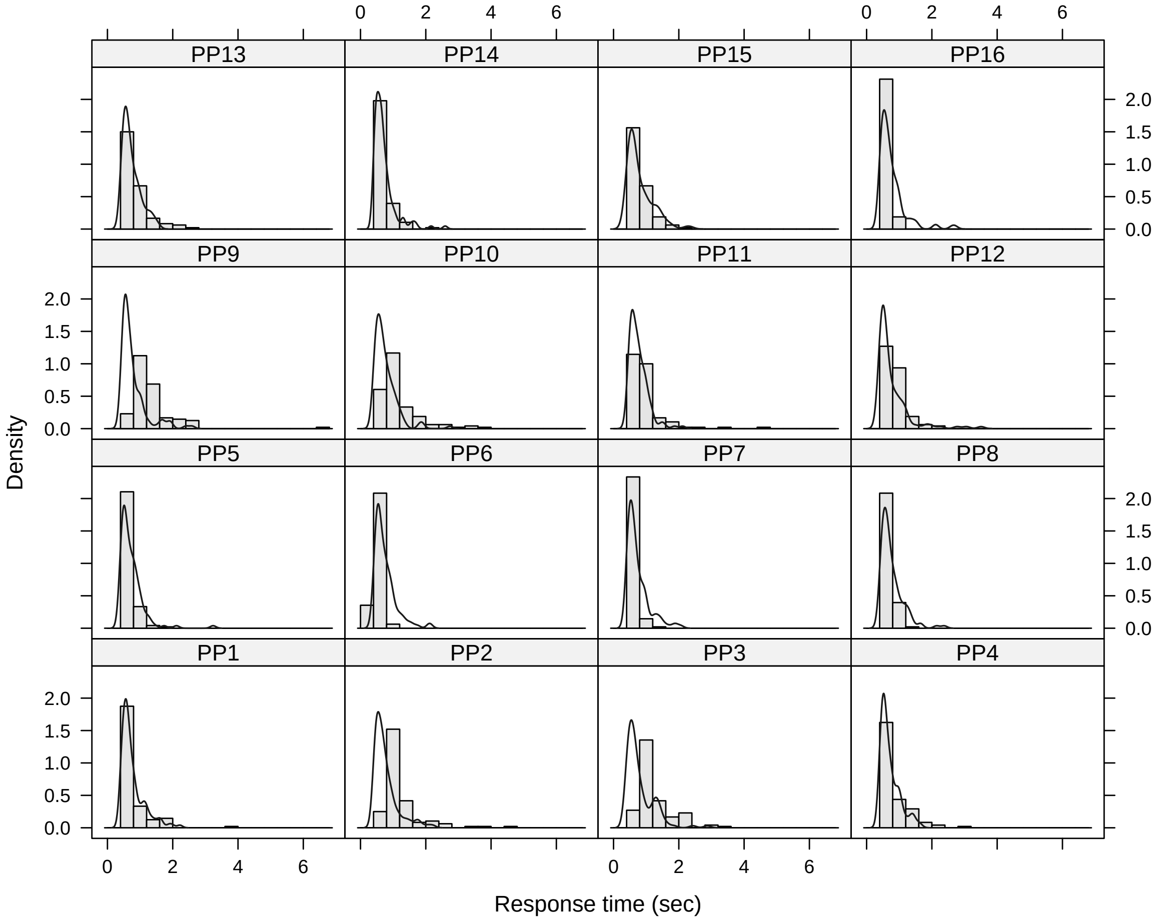


Figure 4.5: Density plots of response time of each participant for the constrained data. The density of the observed RT is displayed via a histogram, and the density of the predicted RT is displayed via a density curve.

to their risk attitudes and decision-making styles. Data of different classifications are generated with different values of the parameters, where risk attitudes are associated with the way subjective values are determined and decision-making styles are associated directly with the level of the decision criterion parameter. The given diffusion model assumes hierarchical structures in constructing priors for the main parameters, which helps immensely to estimate these group-level parameters. Particularly, I use posterior estimates of μ_a and $\boldsymbol{\mu}_u = (\mu_A, \mu_B, \mu_D, \mu_E)$ to estimate the true data-generating group-level parameters for each of the four classifications. The results are summarized in Table 4.5. Note that subjective values are centered with respect to

Gamble C, prior to the simulation, so the current study aims to recover the centered values, not the original values. As seen in Table 4.5, the given model has recovered the group-level parameters to a great precision, with all the true values nicely falling within its 95% credible intervals. For some parameters, it is even possible to locate its true value quite accurately, using just posterior mean.

When it comes to estimating individual-level parameters, I use posterior estimates of a and $\boldsymbol{u} = (u_A, u_B, u_D, u_E)$ to recover the true decision criterion and subjective values for each of the participants (shown in Table 4.2). Figures 4.6 and 4.7 show the results for a and $\boldsymbol{u}$, respectively. Line segments in the figures represent the 95% credible intervals, the empty circled points represent posterior means and the filled circled points represent true values for the respective parameters. All true values are largely consistent with the estimates, where most of the 95% credible intervals include its respective true values within its ranges. Given the results of the simulation study, we can conclude that the given model quite accurately recovers the true data-generating parameters if the data are indeed generated from the same constrained model.

Table 4.5: Estimates for group-level parameters of the constrained diffusion model

Parameter		Classification	True value	Posterior mean	95% credible interval
Decision criterion	μ_a	Conservative	2	2.01	$[1.72, 2.30]$
		Less-conservative	1	1.14	$[0.87, 1.43]$
Subjective values	μ_A	Risk-seeking	2	1.86	$[1.20, 2.54]$
		Risk-averse	-2	-2.14	$[-2.77, -1.47]$
	μ_B	Risk-seeking	1	0.53	$[-0.10, 1.19]$
		Risk-averse	-1	-0.91	$[-1.53, -0.29]$
	μ_D	Risk-seeking	-1	-1.35	$[-2.00, -0.71]$
		Risk-averse	1	0.55	$[-0.08, 1.20]$
	μ_E	Risk-seeking	-2	-2.25	$[-2.93, -1.60]$
		Risk-averse	2	1.69	$[1.04, 2.34]$

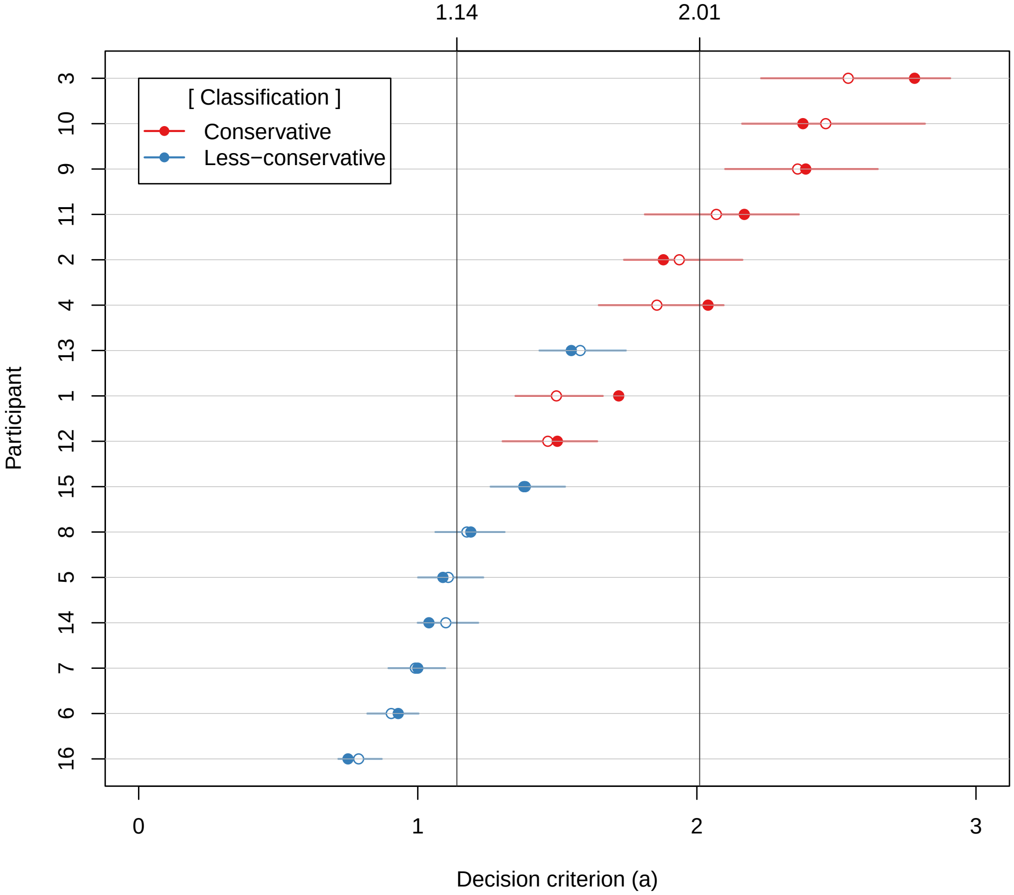


Figure 4.6: Posterior estimates of decision criterion parameter a for the constrained data. Participants are arranged in an ascending order, based on the estimates of a. Empty circled points represent posterior mean, line segments represent the 95% credible intervals, and filled circled points represent true values. Different colors are used to indicate different classifications of decision-making styles, with vertical lines representing the estimates of the group-level decision criterion μ_a for each classification.

4.4.2 Result: The full data

In this section, the constrained diffusion model is fitted to the second data set, where the main data-generating parameters are allowed to vary from trial to trial. The main purpose of this section is, therefore, to see how well the given model is able to recover the true data-generating parameters even when the across-trial variability exists with respect to the main parameters when the data are generated. This is a

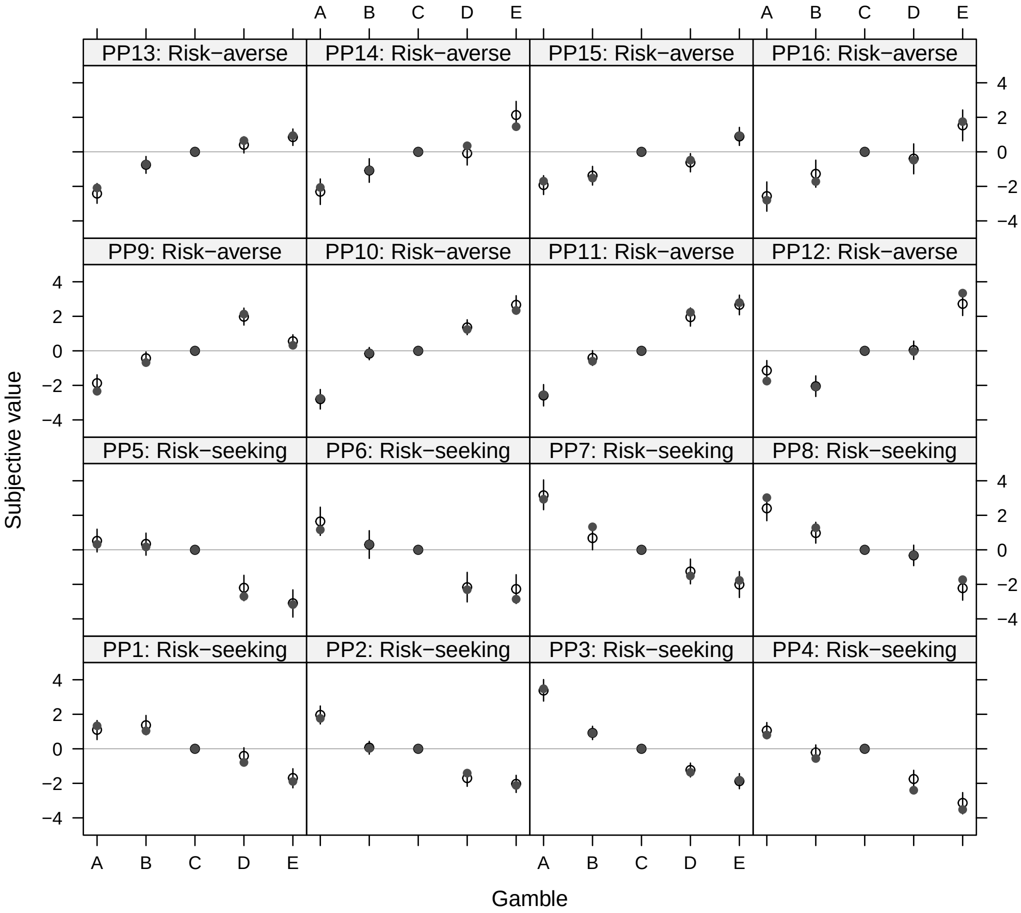


Figure 4.7: Posterior estimates of subjective values $\boldsymbol{u}$ for the constrained data. Empty circled points represent posterior mean, line segments represent the 95% credible intervals, and filled circled points represent true values. Since the subjective values are centered with respect to Gamble C, all subjective values are estimated relative to Gamble C, with Gamble C's subjective value fixed at 0.

tough task for the constrained diffusion model, because the data are generated from a more complicated model within which the current model is nested, yet we still desire to see the same level of performance from the current model as the complicated model would show. Although the results may not be perfect, it is important to learn how far the given model can go even when the setting is not ideal.

As with the analysis of the constrained data set, I begin by evaluating the model fit against the data. As before, I use the posterior samples of the parameters to generate posterior predictive samples $\boldsymbol{y}^{\text{pred}}$, which are, in turn, used to check the model fit. The model fit is evaluated by the Bayesian p-value for choice data and density plots for RT data. The Bayesian p-value for each participant is listed in Table 4.3 under the column named "Full data," and the density plot of whole RT data is depicted in Figure 4.4 as the name of "Full data" as well. Both the Bayesian p-values and the density plot suggest that the constrained diffusion model has no problem accounting for the full data. All the Bayesian p-values are estimated well above 0.05, and the densities of the observed RT (shown as a histogram) and predicted RT (shown as a density curve) show no visible gap between the two, indicating an adequate level of descriptive accuracy provided by the current model for RT data. Individually plotted RT densities (Figure 4.8) also suggest that there are no significant discrepancies between the observed and predicted RT, maybe except for participants 2, 3, and 9. For those participants, the current model tends to predict RT slightly faster than it actually is, with the peak of the density shifted to the left by about half a second. Other than that, the given model seems to provide a decent fit for the data, in terms of both choice and RT data.

Now I examine the main parameters of the diffusion model: decision criterion a and subjective values $\boldsymbol{u}$. As before, I first examine the group-level parameters μ_a and $\boldsymbol{\mu}_u$, then move on to the individual-level parameters. Table 4.6 summarizes the results of the group-level parameter estimation. Posterior mean and the 95% credible intervals are used to estimate the true data-generating parameter values, where posterior mean provides a point estimate and the 95% credible intervals provide a measure of uncertainty as to the estimation, as before.

Table 4.6 shows that the estimates for the main parameters are not quite as accurate as those for the constrained data. Posterior means of the parameters are not near the true data-generating parameter values, and its 95% credible intervals no longer include the true values in its range. However, there is one notable trend in the esti-

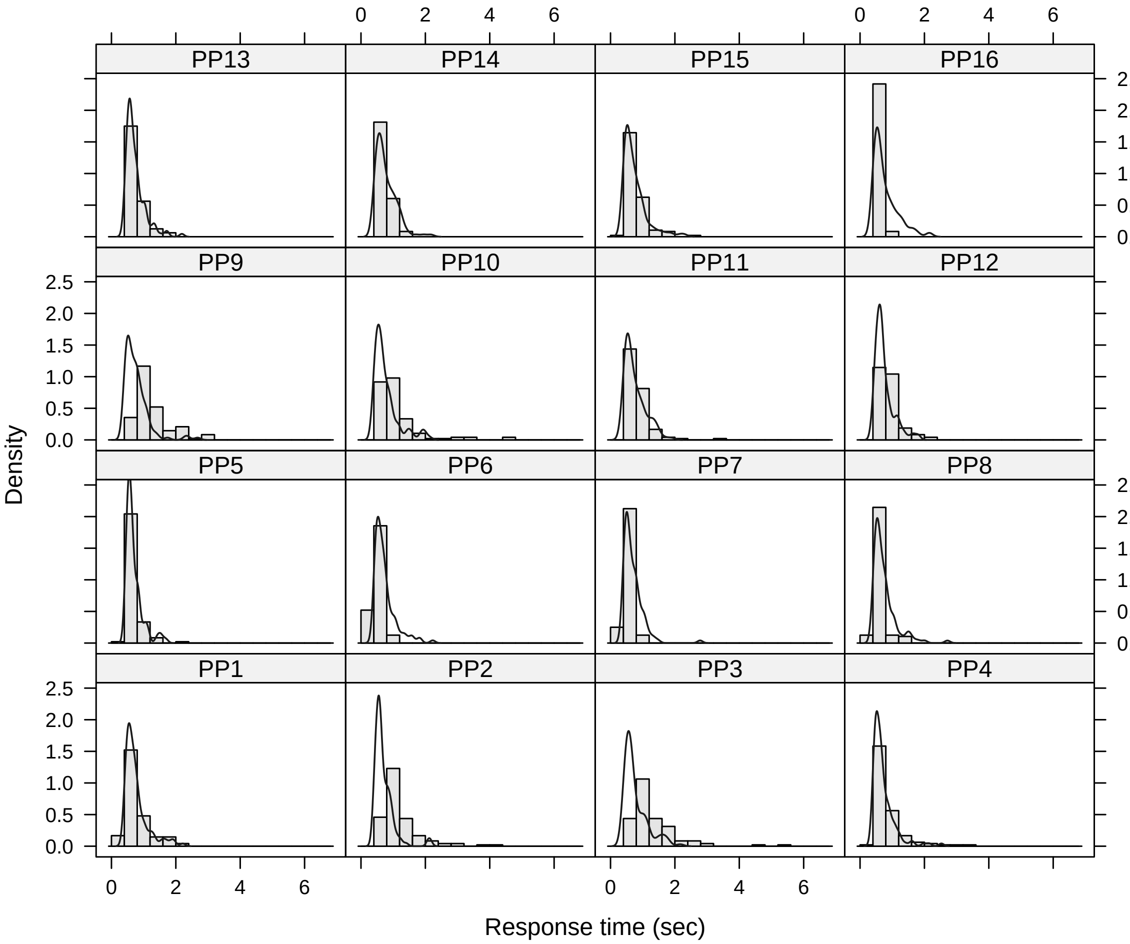


Figure 4.8: Density plots of response time of each participant for the full data. The density of the observed RT is displayed via a histogram, and the density of the predicted RT is displayed via a density curve.

mates: the estimated values are pulled toward its average values. This is a well-known phenomenon of a hierarchical Bayesian model, namely a partial pooling. For example, the decision criterion parameters a for the conservative and less-conservative groups are estimated much closer to their average value, 1.5, compared to the estimates for the constrained data; subjective values $\boldsymbol{u}$ are also all estimated closer to their average value, 0. Such a partial pooling was not obvious for the constrained data, because the constrained data were informative. In other words, the constrained data were generated from the constrained diffusion model, where all across-trial variability is set to 0. Hence, it is straightforward for the constrained model to recover the true values from the constrained data, because every data point was generated from the exact values the model is trying to recover. On the contrary, the full data were generated from the values that vary every trial. Without a capability of accounting for the across-trial variability, the constrained diffusion model is left with no choices but to treat such variability as noise in the data, so more pooling occurs. The partial pooling becomes more evident when we examine individual-level parameters. Figures 4.9 and 4.10 show the results for decision criterion and subjective values. As clear in both figures, individual estimates for each parameter are now much closer to the respective group mean than the ones for the constrained data (Figures 4.6 and 4.7), where more extreme values are likely to suffer a greater decrease in its estimation accuracy due to the partial pooling.

4.5 Discussion

In this chapter, I conduct parameter-recovery simulation studies to test if the constrained diffusion model is able to recover the true data-generating parameter values.

Table 4.6: Estimates for group-level parameters of the full diffusion model

Parameter		Classification	True value	Posterior mean	95% credible interval
Decision criterion	μ_a	Conservative	2	1.74	$[1.58, 1.90]$
		Less-conservative	1	1.21	$[1.06, 1.36]$
Subjective values	μ_A	Risk-seeking	2	1.07	$[0.60, 1.53]$
		Risk-averse	-2	-1.48	$[-1.95, -0.99]$
	μ_B	Risk-seeking	1	0.23	$[-0.23, 0.71]$
		Risk-averse	-1	-0.82	$[-1.26, -0.38]$
	μ_D	Risk-seeking	-1	-0.91	$[-1.37, -0.43]$
		Risk-averse	1	0.57	$[0.10, 1.03]$
	μ_E	Risk-seeking	-2	-1.51	$[-1.98, -1.05]$
		Risk-averse	2	0.92	$[0.46, 1.38]$

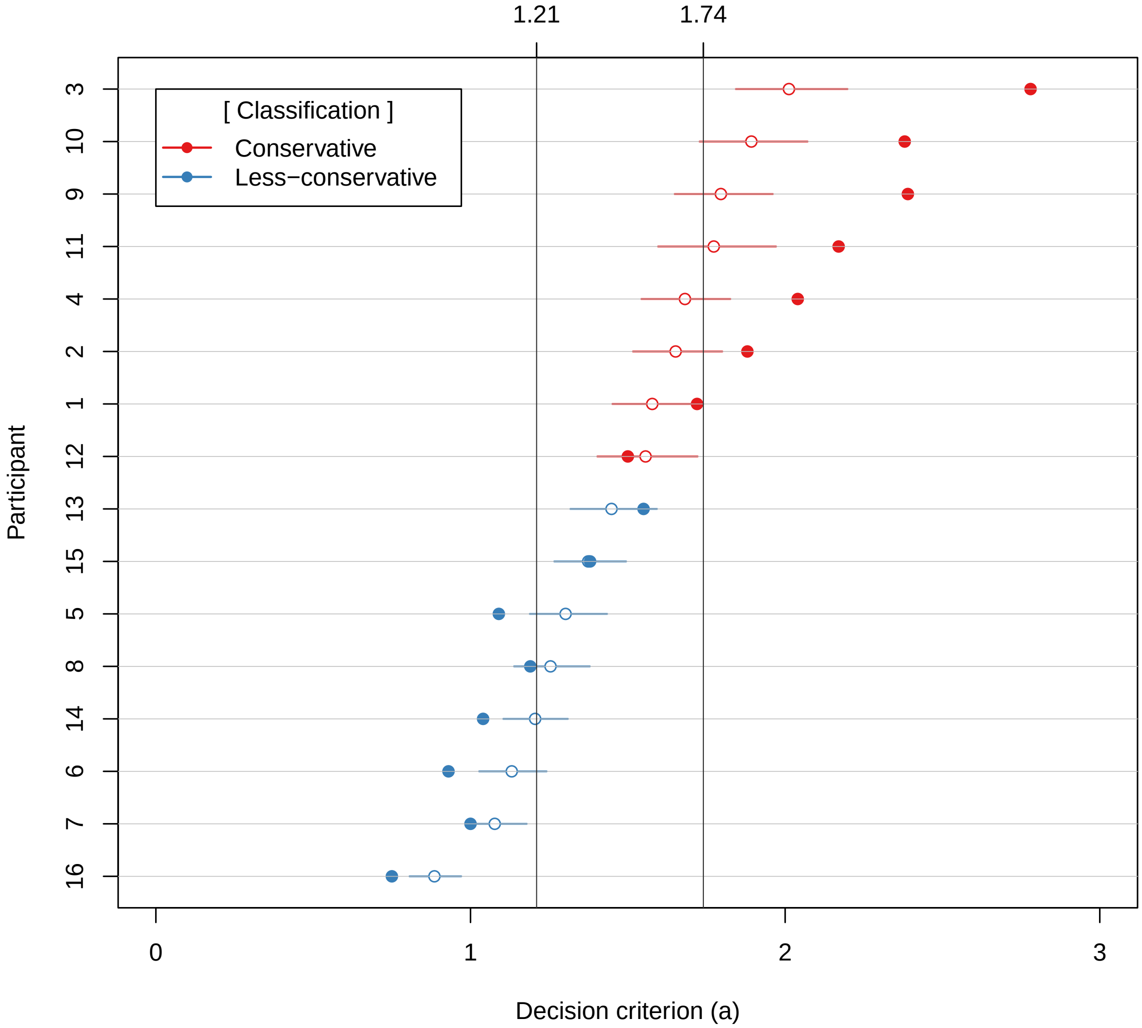


Figure 4.9: Posterior estimates of decision criterion parameter a for the full data. Participants are arranged in an ascending order, based on the estimates of a. Empty circled points represent posterior mean, line segments represent the 95% credible intervals, and filled circled points represent true values. Different colors are used to indicate different classifications of decision-making styles, with vertical lines representing the estimates of the group-level decision criterion μ_a for each classification.

Two data sets are used: one simulated from the same model I choose (called the constrained data) and one simulated from the full diffusion model (called the full data). The given model is able to recover the true values of the parameters to a great precision when the data are generated from the same model I fit the data with. However, the given model is not able to recover the true data-generating values accurately,

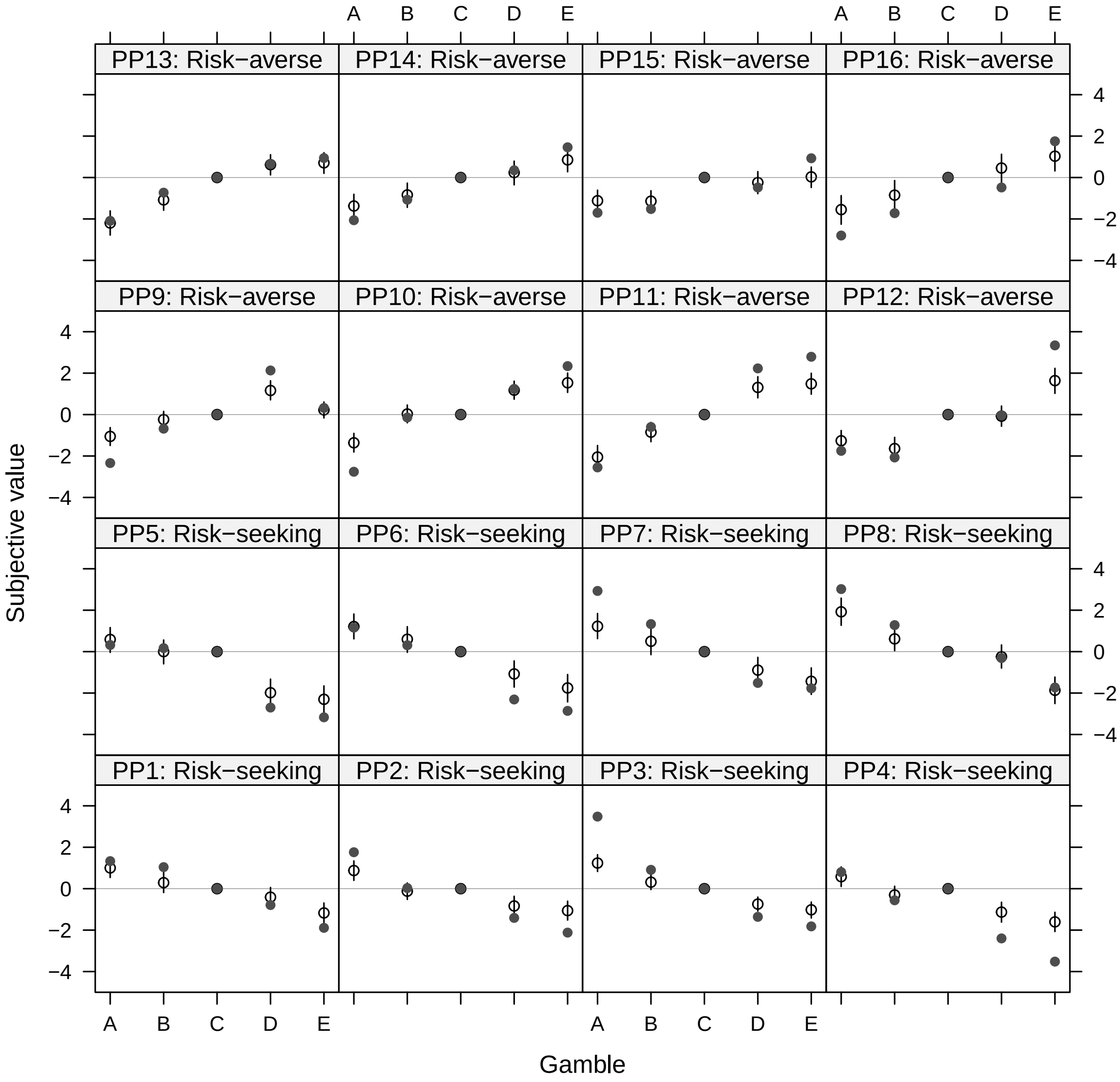


Figure 4.10: Posterior estimates of subjective values $\boldsymbol{u}$ for the full data. Empty circled points represent posterior mean, line segments represent the 95% credible intervals, and filled circled points represent true values. Since the subjective values are centered with respect to Gamble C, all subjective values are estimated relative to Gamble C, with Gamble C's subjective value fixed at 0.

when the across-trial variability of drift rate, decision criterion, and non-decision time are present in the data. Particularly, such variability in parameters has caused the model to generate partial pooled estimates due to its hierarchical structure of priors. The problem is, partial pooled estimates can provide poor estimation accuracy especially when the data are highly noisy, which was evident in the parameter recovery study for the full data. Nevertheless, the current results imply two important points in terms of hypothesis testing.

First, the constrained diffusion model was able to predict which group has higher value of decision criterion for both data sets. Even for the data with across-trial variability in the main parameters, the given model correctly predicted that the conservative group has higher decision criterion than the less-conservative group, where the 95% credible intervals for the two groups are not overlapped with each other. Second, the given model is able to predict the correct rank order of subjective values for both data sets. Although the subjective value estimates were largely off from the true values for the full data set, the predicted rank order within a participant was still preserved. That is, if true subjective values of a participant imply the following rank order, $u_A > u_B > u_C > u_D > u_E$, the estimated values also predict the same rank order, regardless of whether the across-trial variability is assumed or not. These two points are crucial for testing main hypotheses of the current study, and now that the simulation study proves the given model's capability in terms of the two aforementioned points, we can confidently move on to the analysis of real data.

Chapter 5

Empirical analysis: Application to real data

We've discussed the decision theory in Chapter 2, specifically aimed at the axiom of transitivity and lexicographic semiorders, and the diffusion model in Chapter 3. The main goal of the present study is to combine these two great ideas to grasp a better understanding of preferential choices, with particular interest lying in different cognitive processes underneath these two starkly different decision theories, transitivity and lexicographic semiorders. In the literature of preferential choices, transitivity is often considered a pivotal component of one's rationality; if one frequently violates transitivity, there are increasing chances that his or her choices yield sub-optimal results (e.g., money pump; Bar-Hillel & Margalit, 1988), which would question his or her rationality. However, many empirical studies (e.g., Birnbaum & Gutierrez, 2007; Davis-Stober et al., 2015; Regenwetter et al., 2011; Tversky, 1969) have shown that there are almost always at least a few decision makers who routinely violate transitivity. This is where the theory of lexicographic semiorders comes in and describes how we could end up with intransitive preferences. The lexicographic semiorder theory bases its decision-making procedure on a attribute-wise information processing strategy. This strategy shines when the given information is too massive or too complicated to process, by offering a way to circumvent the need of utilizing every piece of information. In other words, the theory produces effective decision-making strategies via the less-is-more principle (Gigerenzer & Brighton, 2009; Gigerenzer & Goldstein, 1996; Gigerenzer & Todd, 1999), that is, by ignoring some pieces of information. The contrast between transitivity and lexicographic semiorders is stark in that transitivity, as mentioned above, is usually described as a quality ascribed to rational decision makers; but lexicographic semiorders are often associated with irrationality, which can embarrass the decision maker himself when he realizes he actually used one of those lexicographic semiorder strategies when he made choices (Tversky, 1969). Thus, the present study aims to examine the underlying cognitive processes of both theories, and by doing so, I seek to provide empirical evidence for different underlying mechanisms of transitivity and lexicographic semiorder, via the lens of the diffusion model.

Note, however, that preferential choices are different from perceptual choices, which the diffusion model (Ratcliff, 1978) had originally targeted. Particularly, in the preferential choice paradigm, the drift rate no longer represents the quality of information about the given stimuli; instead, it represents the difference in subjective value (sometimes referred to as utility) between the given two alternatives (Fudenberg et al., 2018). The subjective value difference usually plays a fundamental role in determining the magnitude of preference in most of the preferential choice studies. Hence, in the current study, I re-interpret and make changes to the original model, so that the model itself can estimate subjective values of individual gambles from the observed data. Therefore, the primary aim of this chapter is to apply the custom built diffusion model to real data sets, all obtained from the preferential choice tasks. Those data sets include data from Cavagnaro and Davis-Stober (2014), data from my own experiment ran in lab, and data from the same experiment but ran online. I start with analyzing the data from Cavagnaro and Davis-Stober's (2014) study.

5.1 Cavagnaro and Davis-Stober's (2014) gamble experiment

5.1.1 Data set

Cavagnaro and Davis-Stober (2014) conducted a set of experiments to test four different probabilistic models of transitivity against data. The experimental design is similar to that of Tversky's (1969), where participants were asked to indicate their preference between two gambles on screen in each trial, but with a few changes. First, the authors added a timed condition to the experiments. In the timed condition, par-

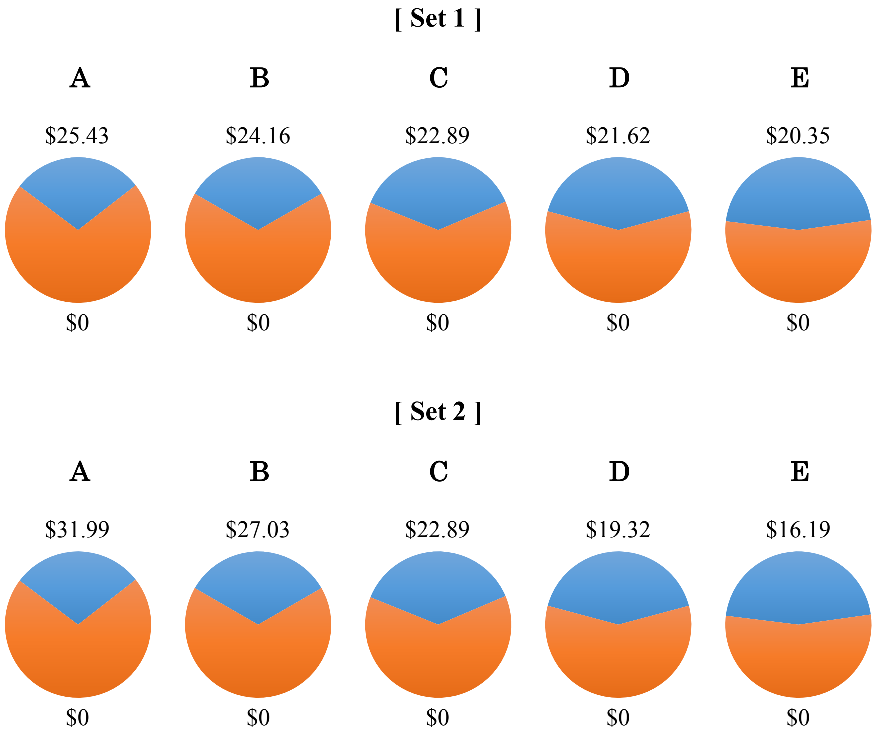


Figure 5.1: Gamble sets 1 and 2 from Cavagnaro and Davis-Stober (2014). Set 1 consists of gambles that have expected value increasing with probability winning. Set 2 consists of gambles that have expected value increasing with payoff.

ticipants were given 4 seconds to make a decision in each trial. If they didn't get to make a decision within 4 seconds, the current trial ended with no response recorded and moved on to the next trial. Second, the authors added a new gamble set, which is shown in Figure 5.1 as Set 2. The original experiment by Tversky (1969) used only one set of gambles, represented as Set 1 in Figure 5.1. The monetary values shown in Figure 5.1 are updated values to reflect the inflation rate since then, with original monetary values being \$4.00, \$4.25, \$4.50, \$4.75, \$5.00 for Gambles A, B, C, D, E, respectively.

The major difference between the two gamble sets, other than their different payoffs, is the different trends of expected monetary values of Gamble Set 1 and Gamble Set 2. Specifically, the gambles in Set 1 have their expected monetary values (given by payoff multiplied by its corresponding probability of winning) increase with probability of winning, while the gambles in Set 2 have their expected monetary values increase with payoff. Table 5.1 shows specifics of both gamble sets. Regardless of which gamble set is used, the basic principle of the experiment remains the same: pairs of gambles that lie far from each other along the chain, such as A and E, generate greater differences in expected value than those that lie close to each other, such as A and B, do.

The experiment allowed no indifference between the given two gambles, so was a 2-alternative forced choice (2AFC) task. The number of participants was 29, and all were students from the University of Missouri. The experiment was designed with five

Table 5.1: Gamble sets

	Gamble Set 1			Gamble Set 2		
Gamble	Prob.	Payoff	Expected value	Prob.	Payoff	Expected value
A	7/24	\$25.43	\$7.42	7/24	\$31.99	\$9.33
B	8/24	\$24.16	\$8.05	8/24	\$27.03	\$9.01
C	9/24	\$22.89	\$8.58	9/24	\$22.89	\$8.58
D	10/24	\$21.62	\$9.01	10/24	\$19.32	\$8.05
E	11/24	\$20.35	\$9.33	11/24	\$16.19	\$7.42

distinct gambles as stimuli, of which pairs of gambles were made. This yielded $\binom{5}{2} = 10$ unique gamble pairs. Each gamble pair showed up repeatedly 12 times throughout the experiment. The experiment was a within-subject design, and so all subjects had to complete all four combinations of the two experimental conditions (i.e., timed/untimed conditions × gamble sets 1 and 2), which makes $12 \times 10 \times 4 = 480$ trials in total per subject. The compensation for participating in the experiment was made after the experimental session. Specifically, the experimenter randomly chose one of the gambles that participants had chosen during the experimental session and let them play it for real money. Participants were told about this way of compensation prior to the main experimental session. This is a typical way of compensating participants in preferential choice tasks, with an attempt to encourage participants to make decisions as if they'd do in the real world. But, this is exactly what made me motivated to design a new experiment on my own, which I discuss in details later.

5.1.2 Results

The goal of the analysis of the data is to examine the difference in the underlying cognitive process between transitivity and lexicographic semiorders. In this analysis, I only analyze the timed data, where all participants were given 4 seconds to respond each trial, because in the un-timed condition, participants frequently made slow responses, which sometimes exceed even 10 seconds. As response time is crucial for the diffusion model analysis and the quality of data directly affects the model's performance with respect to the estimation of the parameters, I decide not to include the un-timed data in the current analysis.

Result: Classification analysis

In this analysis, I classify participants into either transitivity or lexicographic semiorder. As discussed in Chapter 2, there are various ways to model transitivity and lexicographic semiorders, but I choose the Weak Stochastic Transitivity (WST) and the Lexicographic Semiorder Error Model (LSEM) as the main statistical models of transitivity and lexicographic semiorders, respectively, for the current analysis. The reasons are: first, WST provides an intuitive interpretation of the model, along with mathematical ease in its application to the data, and second, WST is one of the most widely applied models for transitivity in the field, so the current analysis could easily fit in the literature. The choice of WST automatically leads to a consideration of the LSEM for lexicographic semiorder, as the LSEM is frequently employed as the lexicographic semiorder's counterpart of the WST (Park et al., 2019).

For the classification analysis, I chose the QTest 2.1 (Zwilling et al., 2019), a public domain software package, which allows for the statistical analysis of order-constrained probabilistic models. Recall that the chosen models, i.e., WST and LSEM, can be recast as a set of order constraints over the space of binary choice probabilities (see Chapter 2 for more discussion). On such models, the QTest 2.1 offers a set of Bayesian inferences, which include Bayesian p-value (Myung et al., 2005), the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002), and Bayes factor (Kass & Raftery, 1995). Of which, I employ the Bayes factor to classify participants' choices into either transitivity or lexicographic semiorders. Bayes factors provide a powerful model selection tool as it quantifies evidence for one model as opposed to another model. Quantifying evidence for a model is clearly an advantage compared to other Bayesian inferences, because Bayesian p-value is only able to tell how poorly the model fits against the data, not speaking for how much the data support the model, and DIC provides a model fit only an ordinal scale, so it makes sense only when the given model is compared to another.

The QTest 2.1 offers three ways to compute the Bayes factor, with all approaches placing the encompassing model in the denominator. So the computed Bayes factors are interpreted as a relative evidence of the model of substantive interest, compared to the encompassing model. By the encompassing model, it means the model that places no restrictions over the space of binary choice probabilities. Thus, the WST and LSEM can be viewed as nested models within the encompassing model, and such nesting features can make the computation of Bayes factors significantly easier (Klugkist & Hoijtink, 2007).

Bayes factors are computed for WST and LSEM and the results are shown in Table 5.3. Since the Bayes factor is a ratio of evidence for the given model over evidence for the encompassing model, a Bayes factor smaller than 1.00 is a sign of the data supporting the encompassing model more than the model of interest, and a Bayes factor greater than 1.00 a sign of the data supporting the model of interest more than the encompassing model. Kass and Raftery (1995) proposed a set of criteria as to how we should make decisions regarding Bayes factors (see Table). They argued that if the model is to be considered a substantially better fit for the given data, as opposed to the competing model, the evidence for the model should be at least twice as much as that for the competing model on the scale of twice the natural logarithm, which provides the same scale of DIC or likelihood ratio test statistics. A Bayes factor of 2 on the twice natural logarithm scale is equivalent to a Bayes factor of 3 on its raw scale, so I employ this criterion to classify the data into WST or LSEM.

However, there is one caveat that may mislead the classification of data: Model

Table 5.2: Decision criteria for Bayes factor

$2\log_e(\mathrm{BF}_{10})$	BF_{10}	Evidence against $\mathcal{M}_0$
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Substantial
6 to 10	20 to 150	Strong
> 10	> 150	Very strong or decisive

mimicry. Note that lexicographic semiorder is not a theory of intransitivity of preferences. Although the lexicographic semiorder theory allows a decision maker to form intransitive preferences, it *can* generate transitive preferences by varying the values of thresholds and order of examination over the given attributes. For example, a decision maker who prioritizes payoff and has a threshold of payoff smaller than the smallest difference in payoff that the given gamble set can make would form a preference of $A \succ B \succ C \succ D \succ E$, which also satisfies transitivity. Thus, by analyzing choice data only, we can't distinguish whether the observed transitive choices have arisen from one of the transitive preferences, or from one of the lexicographic semiorders. For those situations where the Bayes factors favor both models, I classify the data into WST. In other words, for a data set to be classified into LSEM, the data set has to be not classified into WST, but only into LSEM. The third column of each data set in Table 5.3 shows the resulting classification for each participant. And the number of participants classified into either theory is summarized in Table 5.7.

As seen in both Tables 5.3 and 5.4, almost all participants have been classified into either WST or LSEM. Out of 29 participants, only one participant for each gamble set was classified to neither of the two models. Other than that, the two models are quite well able to account for the observed choice patterns, with the dominance of WST. This result is remarkable in my opinion, given that the number of total preference states that can be made out of five gambles is $2^{10} = 1024$, while the number of the preference states that transitivity and lexicographic semiorder predict is only a small fraction of it, i.e., WST predicts 120 different preferences, and LSEM predicts 21 different preferences. Yet, such a small number of preferences can account for the almost entire choice patterns observed in the data. This shows that the two models considered here can make a powerful, yet effective statistical testing tool for the choice data we have.

Now that the classification analysis is done, we move on to the diffusion model

Table 5.3: Bayes factors of Cavagnaro & Davis-Stober (2014)

	Gamble Set 1			Gamble Set 2		
Participant	WST	LSEM	Classification	WST	LSEM	Classification
1	7.26	29.26	WST	1.97	10.19	LSEM
2	3.76	28.24	WST	8.23	23.90	WST
3	0.18	8.26	LSEM	4.02	23.47	WST
4	8.46	40.03	WST	8.40	33.33	WST
5	8.52	47.24	WST	5.45	13.13	WST
6	8.53	48.09	WST	8.50	35.06	WST
7	8.51	45.52	WST	8.53	46.33	WST
8	0.58	16.42	LSEM	0.02	2.66	NA
9	4.67	35.21	WST	7.33	4.22	WST
10	8.02	44.51	WST	8.44	23.08	WST
11	7.95	20.11	WST	8.50	6.24	WST
12	0.00	8.23	LSEM	0.02	10.81	LSEM
13	3.36	8.46	WST	1.05	9.71	LSEM
14	4.46	18.09	WST	1.01	7.47	LSEM
15	3.73	15.59	WST	3.40	1.02	WST
16	8.43	48.17	WST	8.53	41.75	WST
17	8.51	47.57	WST	0.40	3.23	LSEM
18	0.02	41.18	LSEM	3.78	3.15	WST
19	4.14	19.56	WST	5.76	11.58	WST
20	6.70	40.75	WST	5.64	15.51	WST
21	8.50	48.40	WST	8.51	48.48	WST
22	8.51	47.50	WST	4.30	6.60	WST
23	8.53	48.71	WST	8.51	48.63	WST
24	8.53	48.63	WST	8.53	0.08	WST
25	2.27	1.78	NA	1.95	14.52	LSEM
26	0.00	14.12	LSEM	0.00	5.36	LSEM
27	0.20	30.19	LSEM	8.04	27.47	WST
28	6.56	37.58	WST	8.38	26.43	WST
29	7.68	22.43	WST	2.08	5.90	LSEM

analysis with the resulting classifications entering the diffusion model as an explanatory variable for its main parameters. Since the goal of the diffusion model analysis

Table 5.4: Summary of classifications of Cavagnaro & Davis-Stober (2014)

	Number of participants		
Classification	Gamble Set 1	Gamble Set 2	Total
WST	22	20	42
LSEM	6	8	14
NA	1	1	2
Total	29	29	58

is to examine the cognitive processes that underlie transitivity and lexicographic semiorders, I exclude from the diffusion model analysis the participants classified to neither of the two models. Therefore, the number of participants for the diffusion model analysis is 28 for both Gamble Sets 1 and 2, where Participants 25 and 8 are not included in the further analysis of Gamble Sets 1 and 2, respectively. Also, for more reliable results, I set a cutoff value of response time to 0.3 second, so that any response time faster than 0.3 will be considered too fast, and hence eliminated from the analysis as outliers. This cutoff value is particularly due to the guideline suggested by Ratcliff and Tuerlinckx (2002), which addresses how to deal with contaminant response times. Applying the cutoff value of 0.3 seconds to the data, 9 response times (6 from Gamble Set 1 and 3 from Gamble Set 2) are eliminated from the analysis.

Result: Diffusion model analysis

Before I conduct the diffusion model analysis, I first examine response time data for each of the gamble pairs for WST and LSEM. This is an important step that should be taken prior to the application of a model, because an examination of raw data without any modeling assumption may reveal some fundamental characteristics of the true data-generating process that would otherwise be overlooked. Table 5.5 summarizes mean response time for different gamble pairs between WST and LSEM. As seen in Table 5.5, there is a clear tendency in which participants produce slower responses when the two gambles are similar (e.g., A vs B) than when the two gambles are different (e.g., A vs E). Also, there appear to be different patterns in response time between WST and LSEM, at least for some of the data sets, though how the two groups differ is not clear. These findings imply two things: first, people would respond to different gamble pairs differently, so our model should be able to accommodate such different response patterns across different gamble pairs; and second, since the response time patterns appear to be different between WST and LSEM, the classification of each individual should enter the model. The diffusion model I employ for this analysis is particularly designed to reflect those findings, which I briefly go over in the next section.

Model specification For the diffusion model analysis, I use the same model I fitted to the simulated data sets in Chapter 4. The only difference here is the type of the classification that enters the model. In the simulation study, I classified participants according to their decision-making styles (conservative or less-conservative) and risk-attitudes (risk-seeking or risk-averse). The decision-making style of each participant then directly affected the level of the decision criterion parameter a of the diffusion model (conservative participants are associated with higher values of a than less-conservative participants), while the risk attitude of each participant determines the way the gambles are evaluated. In the current analysis, we have the classification of WST and LSEM for each participant. Many decision theories suggest that lexicographic semiorders are featured by its unique decision-making principle, that is, ignoring information. Thus, the classification of WST and LSEM enters the model in the same way as the decision-making styles enter the model in the simulation study. We don't have the information about risk attitudes of the participants, so the group differences are not directly modeled over the mean vector of subjective values. Instead, I hypothesize that whether participants are classified to WST or LSEM affects the variance-covariance matrix of subjective values, because lexicographic semiorders can give rise to cyclic preference patterns, unlike transitivity. Below is the model I fit to the current data.

Let $\boldsymbol{y}_{(pj)}$ be the response vector of person p on trial j, with its components being response $c_{(pj)}$ and response time $t_{(pj)}$. Then, as before, $\boldsymbol{y}_{(pj)}$ are distributed as the Wiener first-passage time distribution (Navarro & Fuss, 2009) (abbreviated to Wiener) as follows:

Table 5.5: Mean response time for gamble pairs between WST and LSEM

		Mean RT (sec) for gamble pairs									
Data set	Classif.	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
CD2014	WST	1.747	1.451	1.424	1.288	1.677	1.483	1.294	1.670	1.480	1.559
Set 1	LSEM	1.778	1.928	1.644	1.618	1.850	1.811	1.721	1.812	1.911	1.886
CD2014	WST	1.728	1.546	1.411	1.333	1.772	1.553	1.492	1.779	1.534	1.628
Set 1	LSEM	1.604	1.686	1.474	1.421	1.600	1.561	1.524	1.554	1.588	1.528
In-lab	WST	1.496	1.420	1.421	1.331	1.613	1.351	1.261	1.501	1.323	1.452
Set 1	LSEM	1.083	0.993	1.026	0.918	1.015	0.907	0.976	1.030	0.982	1.052
In-lab	WST	1.597	1.436	1.350	1.346	1.426	1.389	1.381	1.436	1.326	1.516
Set 2	LSEM	1.073	1.014	1.011	0.994	1.038	1.006	0.947	1.062	1.016	1.086
In-lab	WST	1.563	1.577	1.365	1.438	1.500	1.413	1.314	1.393	1.488	1.376
Set 3	LSEM	1.211	1.006	0.958	0.985	1.194	1.076	1.043	1.097	1.035	1.139
Online	WST	1.266	1.087	0.992	0.955	1.194	1.063	0.987	1.203	1.089	1.122
Set 1	LSEM	1.216	1.162	1.173	1.101	1.201	1.177	1.137	1.176	1.193	1.177
Online	WST	1.144	1.070	1.004	0.966	1.124	1.039	1.028	1.096	1.007	1.095
Set 2	LSEM	1.289	1.253	1.188	1.175	1.273	1.223	1.179	1.275	1.173	1.192
Online	WST	1.180	1.046	0.987	0.950	1.060	1.033	0.983	1.153	0.974	1.049
Set 3	LSEM	1.263	1.226	1.163	1.131	1.236	1.225	1.161	1.210	1.195	1.259

Note. Mean response time tends to be faster for far gamble pairs (e.g., A vs E) than for near gamble pairs (e.g., A vs B). This tendency of response time implies that people in general produce slow responses when the two gambles are similar to each other.

$$\begin{pmatrix} c_{(pj)} \\ t_{(pj)} \end{pmatrix} = \boldsymbol{y}_{(pj)} \sim \text{Wiener}(a_{(p)}, z_{(pj)} = 0.5, t_{er(p)}, v_{(pj)}),$$

where p indexes people, $p = 1, ..., P$, j indexes trials, $j = 1, ..., J$. For each of the main parameters, I consider the following priors. First, for decision criterion a,

$$\begin{gathered} log(a_{(p)}) \sim \text{Normal}(\mu_{a(p)},\ \sigma_a^2),\ \ p \in \{1, ..., P\}, \\ \mu_{a(p)} = \beta_{0a} + \beta_{1a} X_{a(p)}, \\ \beta_{0a} \sim \text{Normal}(0.5,\ 0.5^2), \\ \beta_{1a} \sim \text{Normal}(0,\ 0.5^2), \\ \sigma_a \sim \text{Exp}(0.5), \end{gathered}$$

where $X_{a(p)} = 1$ if participant p is classified to WST, $X_{a(p)} = 0$ otherwise. As before, a log-transformation of a is considered, because it ensures a to be positive. Second, for drift rates v, the current model computes the difference in subjective value between the given gambles. Hence, if the gamble pair A and B is present, for example, drift rate $v_{AB(p)}$ of person p is given by:

$$v_{AB(p)} = u_{A(p)} - u_{B(p)},$$

where $u_{A(p)}$ and $u_{B(p)}$ are subjective values of Gambles A and B for person p. Then, the priors are set for subjective values $\boldsymbol{u}$, not directly for drift rates v:

$$\begin{pmatrix} u_{A(p)} \\ u_{B(p)} \\ u_{D(p)} \\ u_{E(p)} \end{pmatrix} = \boldsymbol{u}_{(p)} \sim \text{MVN}_4\left(\boldsymbol{0},\ \Sigma\right).$$

And I consider the following prior for Σ:

$$\Sigma_{ij} = \eta_k^2 \text{exp}(-\rho_k^2 D_{ij}^2) + \delta_{ij}\sigma_k^2, \;\; i,j \in \{A, B, D, E\}, \; k = 1, 2$$

$$\text{where } \delta_{ij} = \begin{cases} 1 \text{ if } i = j, \\ 0 \text{ if } i \neq j, \end{cases}$$

$$\eta_k^2 \sim \text{Exp}(2),$$

$$\rho_k^2 \sim \text{Exp}(1),$$

$$\sigma_k^2 \sim \text{Exp}(0.5),$$

where k is 1 if the participant is classified to WST, or k is 2 otherwise.

For non-decision time t_{er}, I consider the same prior as in the simulation study as follows:

$$t_{er(p)} = \Phi(\pi_{(p)}) \times \min(\text{rt}_p),$$

$$\pi_{(p)} \sim \;\text{Normal}(\mu_{t_{er}}, \sigma_{t_{er}}^2),$$

$$\mu_{t_{er}} \sim \;\text{Normal}(0, 1),$$

$$\sigma_{t_{er}} \sim \;\text{Exp}(1),$$

Note that the specification of the current model is largely the same as the one in the simulation study. One exception is that as mentioned earlier, we have no information regarding risk attitudes for each participant, so we no longer assume different groups for mean vector of subjective values, which is now replaced by a zero vector. However, the variance-covariance matrix can have different structures according to whether a participant belongs to WST or to LSEM, especially because LSEM can give rise to intransitive preferences. Hence, I add index k, $k = 1, 2$ to the hyper-parameters of Σ, so that the variance-covariance matrix can be estimated separately for each classification.

Posterior predictive check As in the simulation study, I use Stan for a full Bayesian analysis of the specified model against the data. Specifically, four chains are used, with 1,000 posterior samples generated per chain after 1,000 samples of warmup. The information about the MCMC sampling is summarized in Table 5.6.

I examine the results in the same way as in the simulation study: first, I conduct the posterior predictive check of the model, followed by the examination of decision criterion a and subjective values $\boldsymbol{u}$ of the gambles. After the MCMC sampling, convergence of the chains is examined via the number of divergent transitions, number of effective sample sizes for each parameter, Gelman and Rubin's diagnostics (1992) and trace plots. All the diagnostics imply a good convergence of the chains, so I make Bayesian inferences about the model based on the generated 4,000 posterior samples.

First, I check the model fit against the choice data using the Bayesian p-values. As before, I generate posterior predictive samples to compute the Bayesian p-values, which are shown in Table 5.7. Table 5.7 shows that 11 participants have Bayesian

Table 5.6: MCMC sampling information

	Number of	Number of iterations		Averaged time taken (sec)		
Data	chains	Warm-up	Sampling	Warm-up	Sampling	Total
CD2014: Timed Set 1	4	1000	1000	841.06	617.66	1458.72
CD2014: Timed Set 2	4	1000	1000	633.07	601.58	1234.65
In-lab Set 1	4	1000	1000	233.42	141.81	375.23
In-lab Set 2	4	1000	1000	319.70	276.14	595.84
In-lab Set 3	4	1000	1000	260.36	248.10	508.46
Online Set 1	4	1000	1000	3007.15	1432.41	4439.56
Online Set 2	4	1000	1000	2799.30	1387.49	4186.79
Online Set 3	4	1000	1000	2845.53	1470.02	4315.55

Note. The number of chains is set to 4 for all data sets, and the number of iterations per chain is set to 2,000, with 1,000 iterations of warm-up prior to posterior sampling. The time taken to complete the sampling process is averaged across chains and listed in the last three columns.

p-values less than 0.05 for Gamble Set 1, while only 4 participants' p-values are less than 0.05 for Gamble Set 2. Interestingly, most of the participants who have been classified to LSEM end up with p-values less than 0.05 for both gamble sets. This trend becomes more obvious for Gamble Set 1, where all the LSEM-classified participants have p-values less than 0.05. Hence, we see that the current model, or at least some parameter structures of the model, doesn't suit the LSEM data from Cavagnaro and Davis-Stober's experiment quite well. As we will see in the following section, however, the current model provides an adequate fit for most of the participants from the in-lab experiment, even for those who are classified to LSEM. So, I'd argue that the poor performance of the current model with respect to the LSEM choice data is not a global problem, but only limited to the Cavagnaro and Davis-Stober's experiment, thus, I move on to the next results for the moment.

Next, I examine the RT data to evaluate the model fit. As in the simulation study, the model fit with respect to the RT data has been examined via a density plot, where

Table 5.7: Bayesian p-values of Cavagnaro & Davis-Stober (2014)

	Bayesian p-value			Bayesian p-value	
Participant	Gamble Set 1	Gamble Set 2	Participant	Gamble Set 1	Gamble Set 2
1	0.625	0.278	16	*0.025*	0.470
2	0.665	0.695	17	*0.020*	*0.000*
3	*0.000*	0.105	18	*0.000*	0.445
4	0.418	0.328	19	*0.020*	0.490
5	0.807	0.215	20	0.455	0.823
6	0.682	0.578	21	*0.040*	0.360
7	0.672	0.640	22	0.677	0.205
8	*0.000*	NA	23	0.458	0.312
9	*0.002*	0.248	24	0.070	0.268
10	0.085	0.547	25	NA	0.695
11	0.507	0.610	26	*0.000*	*0.000*
12	*0.000*	*0.000*	27	*0.000*	0.290
13	0.340	*0.035*	28	0.262	0.305
14	0.472	0.408	29	0.215	0.128
15	0.060	0.072			

Note. The values less than 0.05 are italicized. For the participants classified to neither of the two models, WST and LSEM, the Bayesian p-values are not computed, but only indicated by NA.

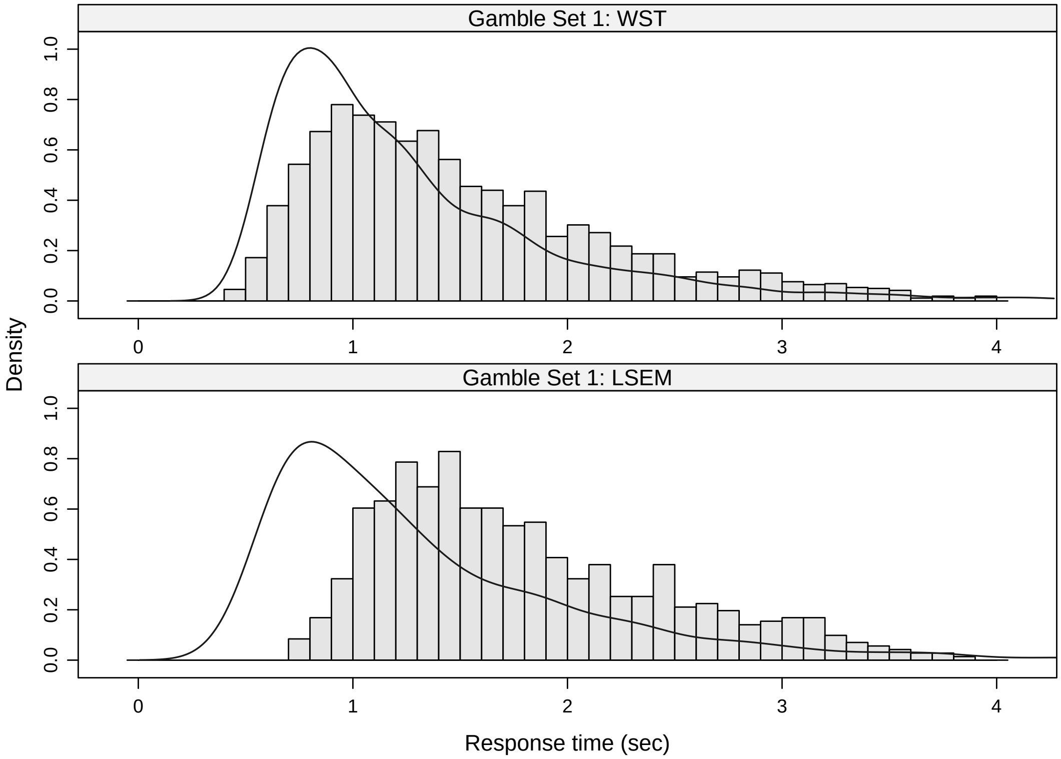


Figure 5.2: Density plot of response time for Gamble Set 1 from Cavagnaro & Davis-Stober (2014). The observed response times are plotted as a histogram, and the predicted response times are plotted as a density curve superimposed on the histogram. Response times are plotted separately for different classifications.

the observed RT data are plotted as a histogram, with the predicted RT data plotted as a density curve superimposed on the histogram. Figures 5.2 and 5.3 show that the model tends to predict response time slightly faster than the observed response time. The gap between the observed and predicted response times grows larger for the LSEM than for the WST, although for Gamble Set 2 shows a bit better fit for LSEM than Gamble Set 1. In sum, the posterior predictive check shows that the given model is underperforming in terms of describing the observed data, and the model's poor performance is specifically associated with the LSEM-classified participants. Hence, we need extra care when interpreting the results of the diffusion model, especially for the LSEM-classified participants.

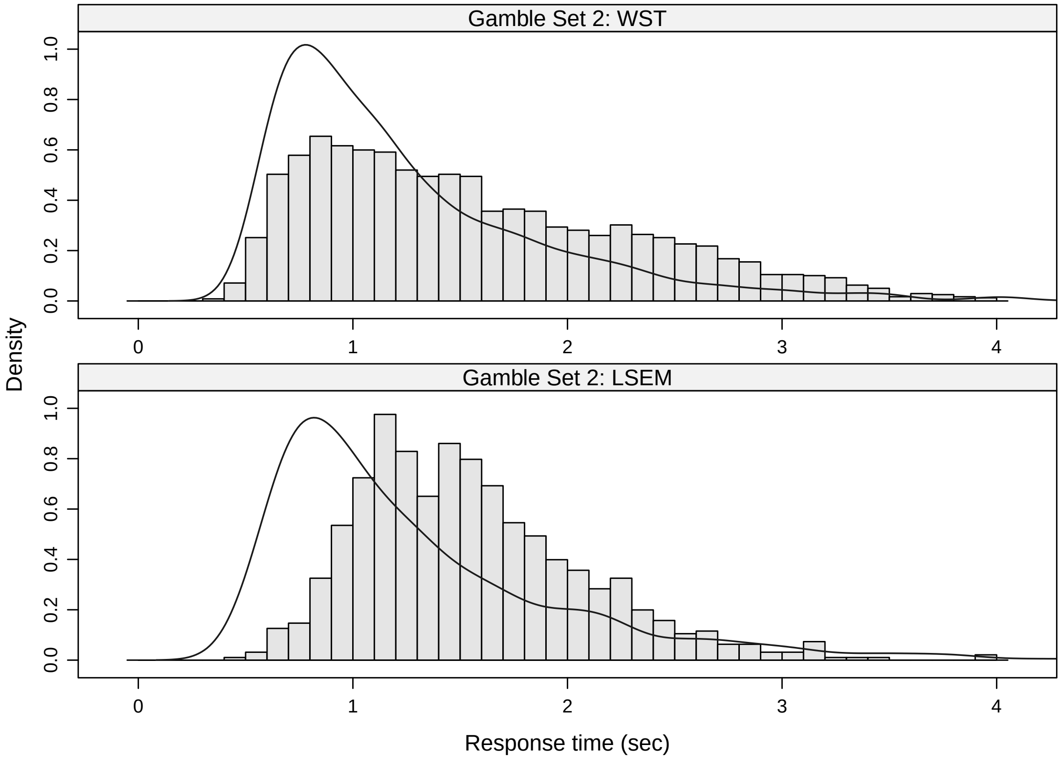


Figure 5.3: Density plot of response time for Gamble Set 2 from Cavagnaro & Davis-Stober (2014). The observed response times are plotted as a histogram, and the predicted response times are plotted as a density curve superimposed on the histogram. Response times are plotted separately for different classifications.

Bayesian inferences on the main parameters As frequently mentioned in this dissertation, the main purpose of the present study is to examine the cognitive processes that underlie transitivity and lexicographic semiorders, via the lens of the diffusion model. And this can be achieved by examining the main parameters of the model, specifically the decision criterion a and subjective values of the gambles $\boldsymbol{u}$. Since the parameters of the diffusion model are directly related to their own cognitive processes, just examining how the main parameters differ between transitivity and lexicographic semiorders provides a good means to fulfill the main purpose of the study.

First, I examine the subjective value parameters $\boldsymbol{u}$ of the gambles across the WST- and LSEM-classified data. Figures 5.4 and 5.5 show the results for each participants,

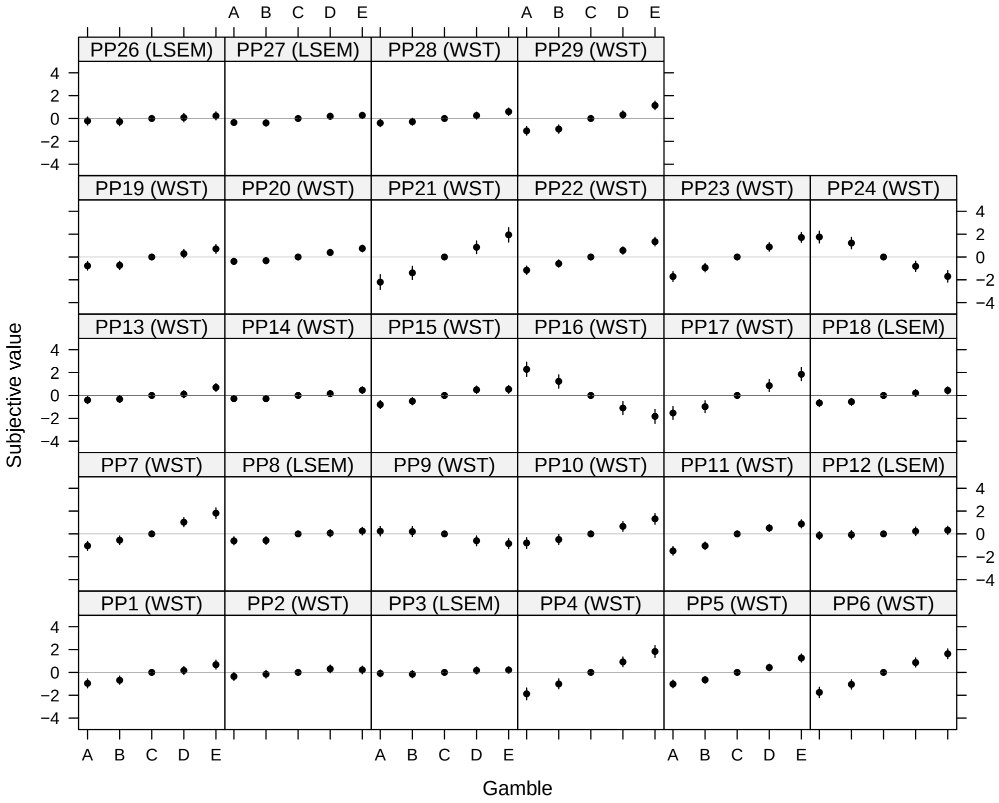


Figure 5.4: Posterior estimates of subjective values for Gamble Set 1 from Cavagnaro & Davis-Stober (2014). Filled dots indicate posterior mean of subjective values of each gamble, and vertical lines overlaid on the filled dots indicate the 95% credible intervals.

where the filled dots indicate the posterior mean of subjective values of the gambles, with the superimposed vertical lines indicating the 95% credible intervals. The resulting classifications are shown next to the participants' id. As the both figures suggest, subjective values of the gambles have different patterns between WST and LSEM. Specifically, the WST-classified participants are likely to have their subjective values linearly ordered across the gambles, while the LSEM-classified participants are likely to have all their subjective values estimated around 0. Such different patterns are deeply related to how each model predicts preferences over the gambles. Recall that the classification in the current analysis is done in a specific way. That is, the participants are classified to WST first, and then, the classification of LSEM is considered only for those who are not classified to WST. This way, we can ensure that no model

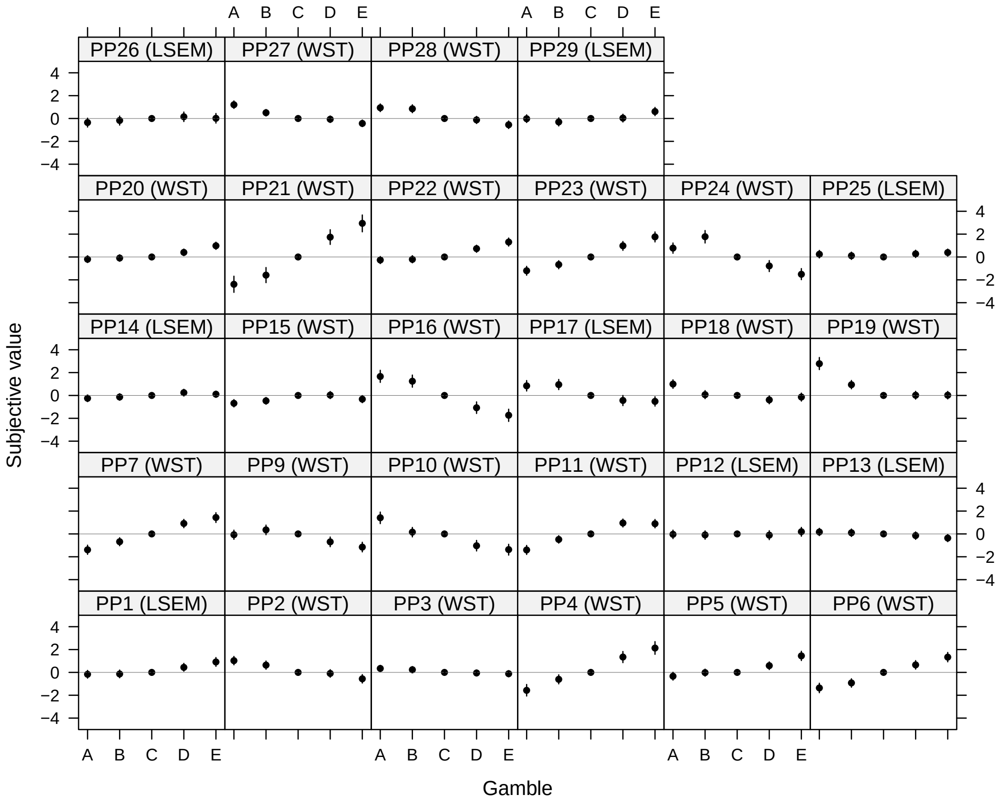


Figure 5.5: Posterior estimates of subjective values for Gamble Set 2 from Cavagnaro & Davis-Stober (2014). Filled dots indicate posterior mean of subjective values of each gamble, and vertical lines overlaid on the filled dots indicate the 95% credible intervals.

mimicry exists in those classifications, and at the same time, we can make sure that lexicographic semiorder favors only the participants who have intransitive preferences over the gambles. Since intransitive preferences feature cyclical orders of preferences, all subjective values are estimated around 0 for the LSEM participants. That is, the model couldn't decide which gamble the participants value higher than others, due to the cyclical patterns of preferences.

Second, I examine the decision criterion parameter a of the WST- and LSEM-classified data. Chief interest lies in the difference in a between WST and LSEM, and I evaluate whether the estimated difference is statistically significant by using the posterior probability. This is one of the widely suggested methods for testing hypothesis the Bayesian way (Gelman et al., 2013; Jackman, 2009), along with Bayes factor, Bayesian p-value, and DIC, where the posterior density is used to compute the posterior probability of a test statistics of interest. In the current analysis, I am interested in testing whether the decision criterion of WST is higher than that of LSEM, so the test statistic will simply be the difference in a between the two groups. Note, however, that a is individually estimated for each participant, so if we aim to compare a between groups, we need to compare group mean of a, i.e., μ_a, between WST and LSEM. Hence, the test statistic of interest is $\mu_a^{\text{WST}} - \mu_a^{\text{LSEM}}$, and it can be easily obtained by the posterior samples of μ_a^{WST} and μ_a^{LSEM}. The results are shown in Table 5.8.

Posterior estimates of μ_a^{WST} and μ_a^{LSEM} for Cavagnaro and Daivs-Stober's (2014) data, which are shown in first two rows of Table 5.8, imply that participants of WST tend to have higher values of a than those of LSEM. This tendency is evident in Gamble Set 2, where majority of the posterior density of $\mu_a^{\text{WST}} - \mu_a^{\text{LSEM}}$ is greater than 0. Specifically, $p(\mu_a^{\text{WST}} - \mu_a^{\text{LSEM}} > 0|y) = 0.995$ for Gamble Set 2. This result provides a theoretically consistent account of lexicographic semiorders, which characterize itself under the "less-is-more" principle, that is, the feature of ignoring information. I would like to stress the importance of such results, because two different theories with two supposedly different cognitive processes are now supported by empirical evidence.

Table 5.8: Posterior estimates of μ_a^{WST} and μ_a^{LSEM}

Data Set	μ_a^{WST}	μ_a^{LSEM}	$p(\mu_a^{\text{WST}} - \mu_a^{\text{LSEM}} > 0\|\boldsymbol{y})$
CD2014 Set 1	2.25 [2.05, 2.44]	2.08 [1.71, 2.43]	0.793
CD2014 Set 2	2.36 [2.19, 2.55]	1.94 [1.68, 2.20]	0.995
In-lab Set 1	2.13 [1.91, 2.34]	1.63 [1.39, 1.87]	0.997
In-lab Set 2	2.08 [1.77, 2.41]	1.74 [1.35, 2.12]	0.920
In-lab Set 3	2.11 [1.75, 2.45]	1.77 [1.44, 2.09]	0.932
Online Set 1	1.94 [1.83, 2.04]	1.89 [1.80, 1.98]	0.748
Online Set 2	1.90 [1.79, 2.00]	1.95 [1.86, 2.05]	0.219
Online Set 3	1.89 [1.79, 2.01]	1.96 [1.87, 2.05]	0.191

Note. Posterior mean of μ_a^{WST} and μ_a^{LSEM} are listed, with its 95% credible intervals placed in between square brackets. Posterior probability of μ_a^{WST} being greater than μ_a^{LSEM} is also computed and listed in the last column

Particularly, lexicographic semiorders are known to ease the decision-making process by ignoring information, and the current diffusion model analysis shows that the exact parameter that governs the amount of information needed to make a decision is the one that differs significantly between WST and LSEM. It's always exciting to see that the two different theories speak to the same result even though the two theories come from different disciplines. The results all indicate that participants of LSEM would require significantly less amount of information to make a decision compared to those of WST.

However, the result of Gamble Set 1 doesn't provide strong evidence for such a tendency as in Gamble Set 2. I suspect that the way the participants were compensated for their participation may have something to do with this result. The participants in Cavagnaro and Davis-Stober's (2014) experiment were informed that they would play one of the gambles they choose for real money at the end of the experiment. This information was given prior to the main experimental session, so I believe that participants' choices were affected by this information. In other words, the participants would be biased toward the gambles with higher probability of winning, because they would like to increase the chance of winning real money. Consider this tendency in the context of different gamble sets. As shown in Figure 5.1, expected values of Gambles in Set 1 increase with probability of winning, while expected values of Gambles in Set 2 increase with payoff. Thus, Gamble Set 1 provides expected values consistent with the tendency to choose higher probability of winning, whereas Gamble Set 2 provides expected values in conflict with the tendency for higher probability of winning. As we discussed before, transitivity and lexicographic semiorders differ the most when conflicts of information are present, so I think that's the reason why we see strong empirical evidence in Gamble Set 2, not in Gamble Set 1. And this is exactly what motivates me to design a new experiment, in the following section.

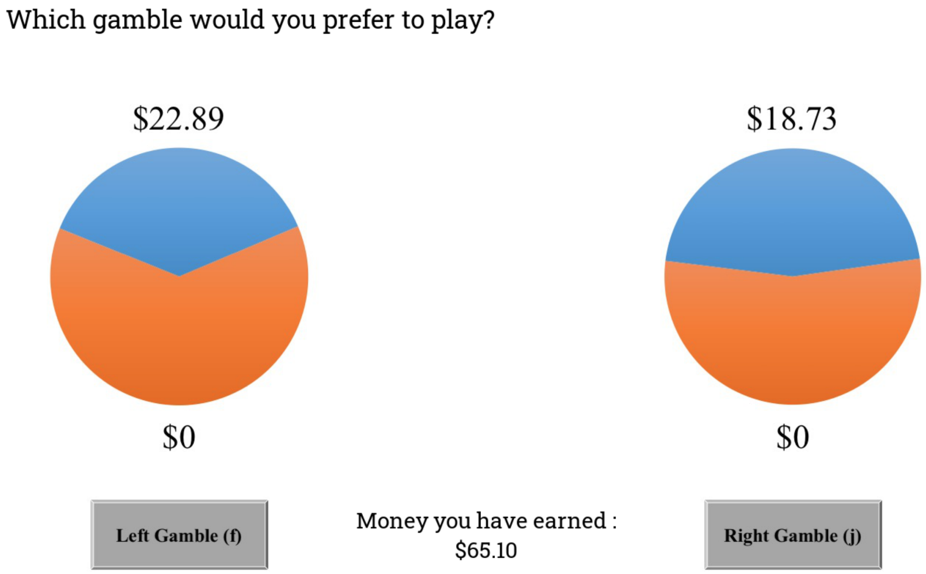
5.2 In-lab experiment

5.2.1 Data set

As I elaborate above, I question the way the previous experiment compensated participants for their participation. Such a way of compensating participants as in Cavagnaro and Daivs-Stober's (2014) experiment is a typical protocol, widely practiced especially in economic preferential choice experiments. In those experiments, the experimenter would inform the participants that after the experiment, they are going to play one of the gambles they choose during the experimental session for real money. Such information is usually provided before the main experimental session begins, because most researchers thought that this way of compensation would encourage participants to make decisions as if they would actually do in the real world. However, I question the validity of this approach for compensating participants for the following reasons. First, unlike the real world, participants don't get to receive feedback for every choice they make in the experiment. Even if participants get to play one of the gambles they choose during the experiment, that's just one delayed feedback of their choices, not able to test their choices against the its outcomes every trial. In this case, participants might just go for the specific gambles without much thinking about it, especially when participants consider the task repetitive and tedious. Second, since the compensation is given only when participants win the chosen gamble, the participants might form a tendency toward probability of winning during the experiment. This is not a real tendency of participants, but rather an artificial one, I say, created by the experimental design. For these reasons, I decided to design my own experiment, which addresses the points just mentioned here. [1]

I keep the basic paradigm same as the previous experiments, but incorporate

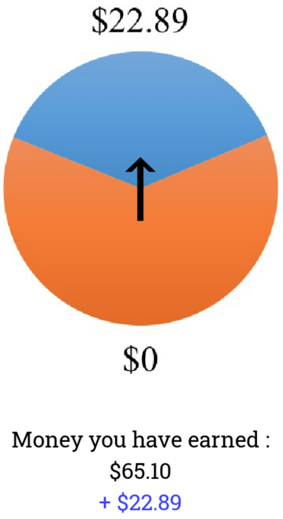
[1] I create this new experiment using Python. Specifically, I use PyGame, a cross-platform set of Python modules originally designed for writing video games.



Which gamble would you prefer to play?

Money you have earned :
$65.10

Figure 5.6: An example trial of the new experiment. Now participants get to play the chosen gamble every trial, and the money they win is accumulated in their virtual account. How much money they have earned is shown in the middle on screen.



Money you have earned :
$65.10
+ $22.89

Figure 5.7: An example of the chosen gamble being played. If the arrow points to the blue-colored area, participants win the specified amount of money. Otherwise, they win nothing.

one major difference into the new experiment: participants get to play the chosen gamble after each trial. Every trial participants are asked to choose one from a pair of gambles. Once the decision is made, they now get to play that chosen gamble for virtual money. If they win, the money they win is saved in their virtual account, which is shown in the middle of screen. In other words, every trial, participants would see a pair of gambles, along with the total amount of money they have earned so far. Figure 5.6 shows an example trial of the new experiment, and Figure 5.7 shows how the chosen gamble is played. As shown in Figure 5.7, when a gamble is being played, an arrow appears in the middle of the chosen gamble. If the arrow points to the blue-colored area, participants win the specified amount of money, and the earned money goes to the virtual account shown in the middle of screen. If the arrow points to the red-colored area, participants win nothing, so nothing happens to the virtual account and the amount of money they've earned remains the same that trial. In this experiment, participants are encouraged to earn as much money as possible, which, in my opinion, provides a good motivation for making decisions that emulate the way they actually make decisions in the real world, and also effectively eliminates the possibility of being biased toward a specific aspect of gambles.

Also, I add Gamble Set 3 to the experiment (see Figure 5.8). The gambles in Set 3 share the same expected monetary value with each other, so it will provide a nice middle ground between Sets 1 and 2. The rest of the experimental design remains the same as before. The experiment is the 2AFC task, so no indifference between the given two gambles is allowed. The number of participants is 12, and the number of repetition of each gamble pair is set to 12. The experiment is a within-subject design, so each participant had to complete all three gamble sets, which results in $3 \times \binom{5}{2} \times 12 = 360$ trials total. Participants were given 4 seconds to make a decision every trial as in the timed condition in Cavagnaro and Davis-Stober's (2014) experiment. The average time taken to complete the whole experimental session per participant was about 35 minutes.

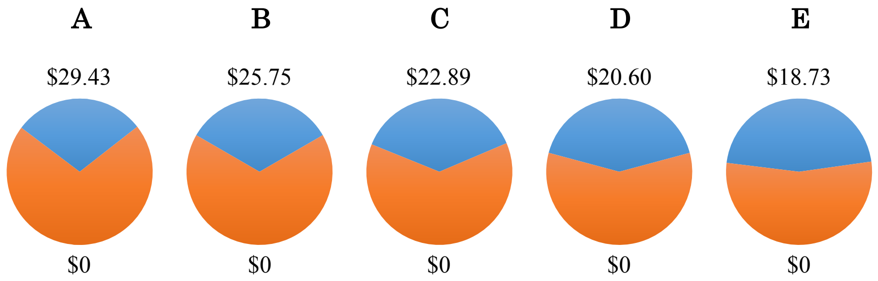


Figure 5.8: Gamble Set 3. All the gambles in Set 3 have the same expected monetary value.

5.2.2 Results

Classification analysis

As in the previous section, I first conduct the classification analysis using the Bayes factor. Specifically, I classify participants into either WST or LSEM according to their Bayes factors, with Kass and Raftery's (1995) decision criteria (see Table 5.2). Hence, any Bayes factor greater than 3 is considered substantial evidence for the given model over the encompassing model. Again, I use the QTest 2.1 (Zwilling et al., 2019) to compute the Bayes factors. The results are shown in Tables 5.9 and 5.10. For classifications, I first attempt to classify participants into WST, and then I attempt to classify only those who are not classified to WST into LSEM. This is due to the model mimicry, so this way of classifying participants ensures that only the WST-classified participants have transitive preferences, which automatically leads the LSEM-classified participants to have intransitive preferences.

As shown in Tables 5.9 and 5.10, the two models, WST and LSEM, are well able to account for almost all the choice patterns observed in this data set, leaving only one participant for Set 3 unclassified. Again, this is a remarkable result given that the two models can account for only a small number of preference states, compared to the total number of preference states allowed for participants. One thing worth noting in this result is the distribution of WST and LSEM classifications. Unlike the data set

Table 5.9: Bayes factors of in-lab data

	Gamble Set 1			Gamble Set 2			Gamble Set 3		
PP.	WST	LSEM	Classif.	WST	LSEM	Classif.	WST	LSEM	Classif.
1	3.30	1.87	WST	4.20	14.23	WST	3.81	18.98	WST
2	4.76	5.12	WST	1.38	10.82	LSEM	1.91	4.52	LSEM
3	1.68	13.85	LSEM	1.78	17.32	LSEM	0.72	1.57	NA
4	8.49	48.39	WST	3.94	0.13	WST	7.95	0.52	WST
5	0.34	34.93	LSEM	0.63	18.67	LSEM	0.69	28.26	LSEM
6	6.44	13.49	WST	3.66	13.66	WST	2.48	7.93	LSEM
7	5.03	8.88	WST	3.78	9.87	WST	3.46	16.95	WST
8	1.60	13.80	LSEM	4.79	15.13	WST	1.77	13.62	LSEM
9	7.50	42.73	WST	6.81	44.73	WST	7.03	45.67	WST
10	2.35	19.47	LSEM	1.98	20.78	LSEM	1.46	9.18	LSEM
11	0.25	5.19	LSEM	0.60	29.20	LSEM	0.85	7.84	LSEM
12	7.81	37.05	WST	5.56	0.44	WST	4.66	1.90	WST

from Cavagnaro and Davis-Stober (2014), the current data set shows an almost evenly distributed classifications between WST and LSEM. This is an interesting pattern of the classifications, because numerous research has been reporting the dominance of transitivity among the data (e.g. Birnbaum & Gutierrez, 2007; Cavagnaro & Davis-Stober, 2014; Davis-Stober et al., 2015). Perhaps it's the experimental design I newly employed, particularly the one, where participants get to play the chosen gamble every trial, that may have affected the distribution of the classifications. The reason is that in the newly designed experiment, participants not only have to consider whether the choices they make are consistent with their preferences, but also do they need to consider the results of a gamble play in the previous trials to make better choices. In other words, there are more factors for participants to consider in this experiment,

Table 5.10: Summary of classifications of in-lab data

	Number of participants			
Classification	Gamble Set 1	Gamble Set 2	Gamble Set 3	Total
WST	7	7	5	19
LSEM	5	5	6	16
NA	0	0	1	1
Total	12	12	12	36

which may have affected the distribution of WST and LSEM.

Regardless of the reason, evenly distributed groups make a good candidate for a group variable of a model. The resulting classification again enters the diffusion model as a group variable for the decision criterion parameter a. We now turn to the diffusion model analysis.

Diffusion model analysis

I use Stan for a full Bayesian analysis of the diffusion model. The number of the chains was set to 4, and the number of iterations for posterior sampling was set to 1,000 per chain (see Table 5.6). As before, I exclude responses faster than 0.3 second from the diffusion model analysis for reliable results. As a result, 60 responses are excluded from the analysis as outliers (i.e., 25 responses from Set 1, 16 responses from Set 2, and 19 responses from Set 3 are excluded). Convergence of the chains is confirmed by the number of divergent transitions, number of effective sample size for each parameter, Gelman and Rubin's diagnostics, and trace plots. All the diagnostics indicate that the chains have been mixed well, and converged to the target density, that is, the joint posterior density.

Posterior predictive check As before, I use the posterior predictive samples to compute the Bayesian p-values to evaluate the model fit against the choice data. And I examine the density plots of the observed and predicted RT to evaluate the model fit against the RT data. Table 5.11 shows the computed Bayesian p-values.

The first thing that comes to view is the much improved model fit of the diffusion model against the observed choice data. Compared to the data set from Cavagnaro and Davis-Stober (2014), the Bayesian p-values in Table 5.11 show much better fit for the data, with only two participants (Participants 4 and 9) not fitted well by the model across different gamble sets. Density plots of the observed RT (plotted as a histogram) and predicted RT (plotted as a density curve) also show a good overlap with each other (see Figures 5.9, 5.10, and 5.11). One notable trend we see from the density plots is that the LSEM-classified participants are noticeably faster than the WST-classified participants. A closer look at the plots reveals that the histograms of the observed RT for the LSEM appear to have a peak around half a second and another peak around 1 second. The first peak tends to have lower density than the second one, yet still substantial amount of density is located around the first peak, while the second peak of the LSEM data generally matches the peak of the WST data. It seems as though there exists a common cognitive process that underlies both WST- and LSEM-classified participants, with an additional underlying process uniquely for the LSEM-classified participants. Since the current diffusion model can't account for such bimodality of the RT distribution, we might need to consider different models if we are interested in modeling the observed RT in a rigorous manner.

Bayesian inferences on the main parameters As in the analysis of Cavagnaro and Davis-Stober's (2014) data in the previous section, the main interest of

Table 5.11: Bayesian p-values of in-lab data

	Bayesian p-value		
Participant	Gamble Set 1	Gamble Set 2	Gamble Set 3
1	0.605	0.210	0.642
2	0.720	0.128	0.470
3	0.350	0.613	NA
4	0.090	*0.000*	*0.007*
5	0.445	0.482	0.530
6	0.185	0.230	0.462
7	0.172	0.298	0.115
8	0.165	0.510	0.375
9	*0.005*	*0.000*	*0.002*
10	0.595	0.573	0.182
11	0.098	0.375	0.202
12	0.465	0.560	0.160

Note. p-values less than 0.05 are italicized, which indicate lack-of-fit of the model against the data.

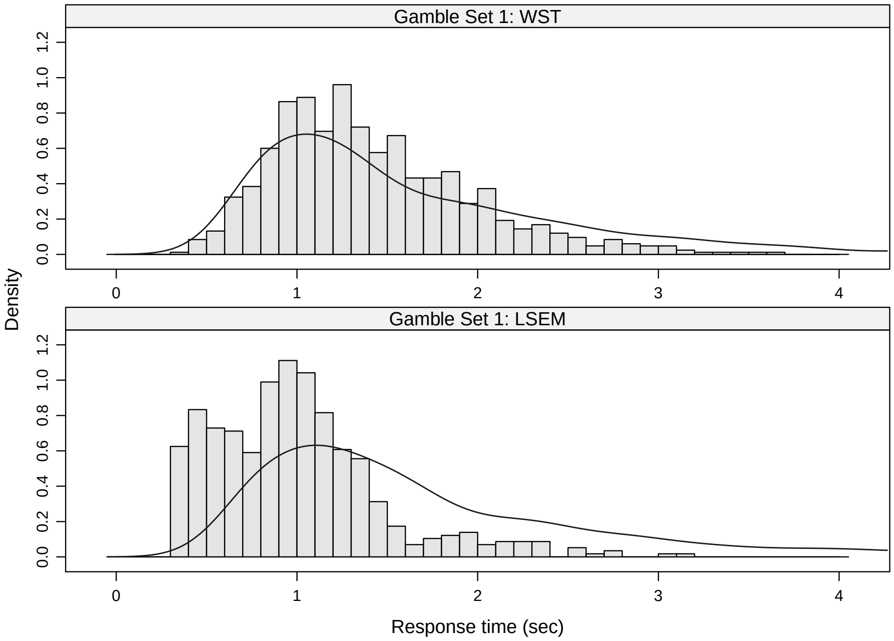


Figure 5.9: Density plots of RT for Gamble Set 1 in in-lab data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve.

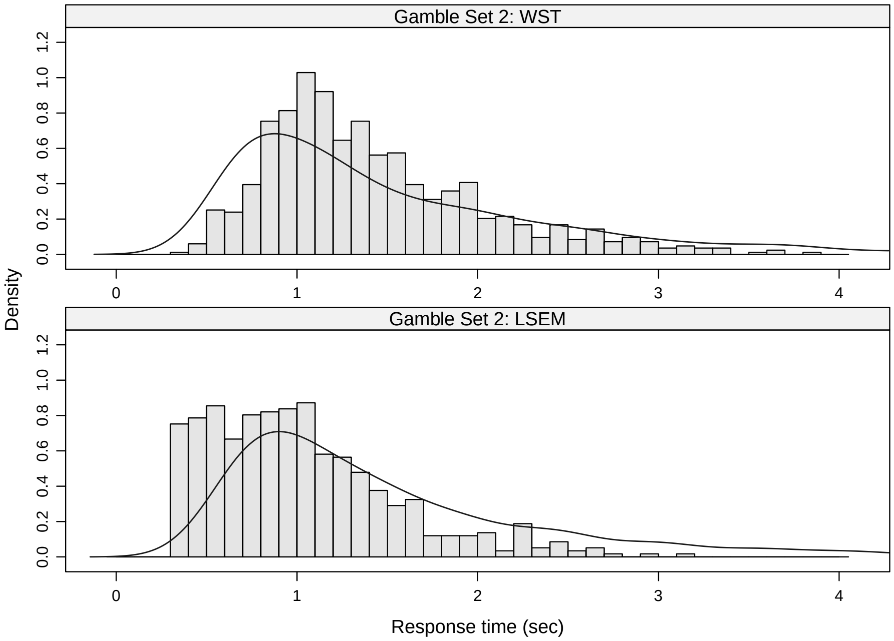


Figure 5.10: Density plots of RT for Gamble Set 2 in in-lab data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve.

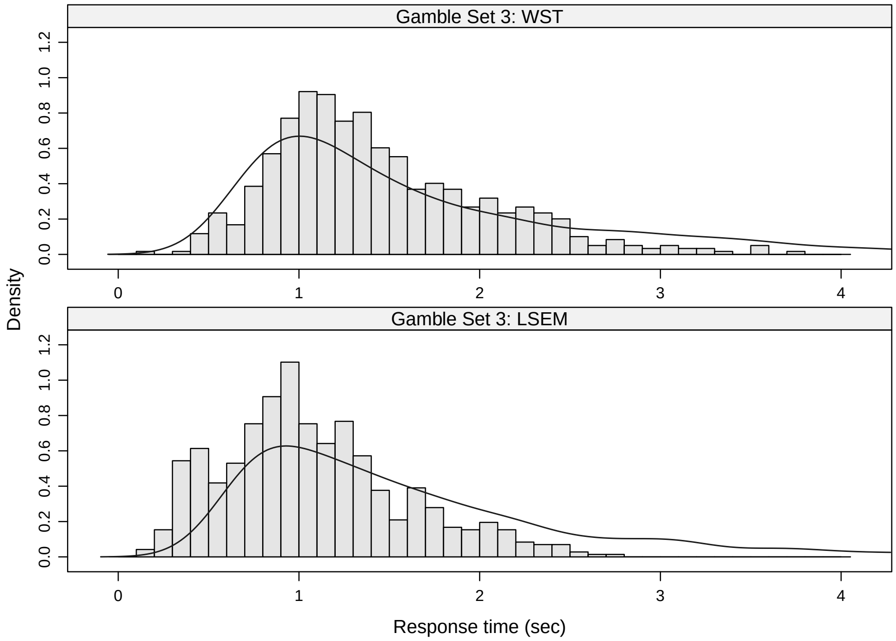


Figure 5.11: Density plots of RT for Gamble Set 3 in in-lab data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve.

the diffusion model analysis lies in the decision criterion a and subjective values $\boldsymbol{u}$. I examine these parameters in the same way as I did for Cavagnaro and Davis-Stober's (2014) data; that is, I examine posterior estimates of group mean of a for each classification, μ_a^{WST} and μ_a^{LSEM}, and the trend of the estimated subjective values for each classification. Table 5.8 shows the posterior estimates of μ_a^{WST} and μ_a^{LSEM} and Figures 5.12, 5.13, and 5.14 show estimated subjective values of the gambles.

All the results from the analysis of the in-lab data support the main idea of the current analysis: lexicographic semiorders integrate less information to produce a decision, and its underlying preferences form cyclic patterns. In Table 5.8, the group mean of a for the WST, μ_a^{WST}, is significantly greater than that for the LSEM, μ_a^{LSEM}, across all gamble sets, with its posterior probabilities of μ_a^{WST} being greater than μ_a^{LSEM} equal to .997, .920, and .932 for Gamble Sets 1, 2, and 3, respectively. And subjective values shown in Figures 5.12, 5.13, and 5.14 imply cyclic preferences for the LSEM, because all the subjective values are estimated around 0, as no one

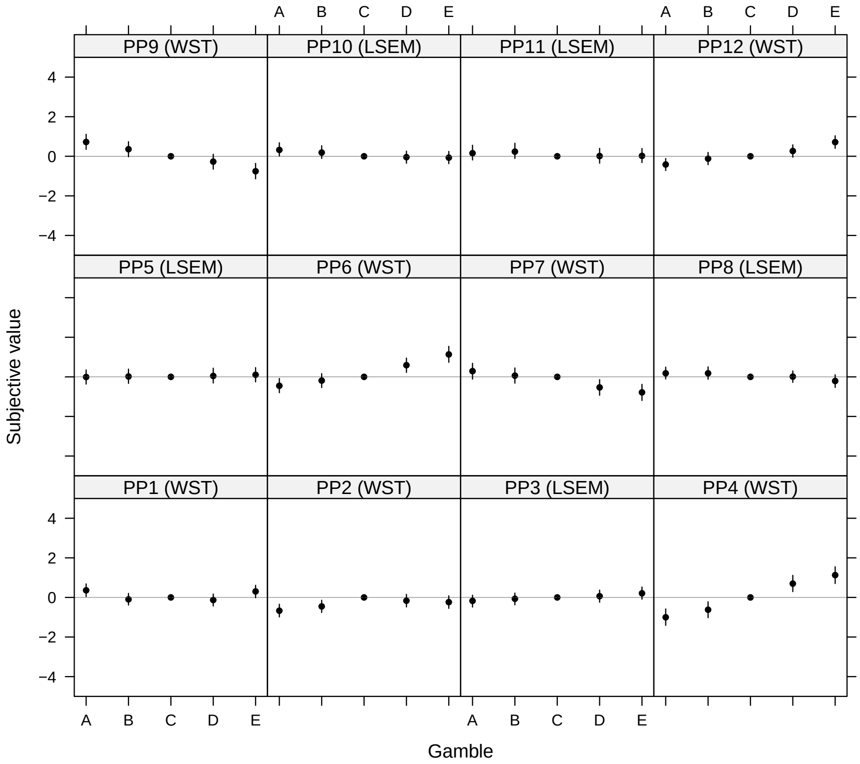


Figure 5.12: Subjective values of Gambles in Set 1 in in-lab data. Filled dots indicate posterior mean, and vertical lines superimposed on the dots indicate the 95% credible intervals.

subjective value stands out due to its cyclic pattern of preferences.

5.3 Online experiment

5.3.1 Data set

The same experiment was administered online, primarily due to the COVID-19. I use PsyToolkit to administer the online experiment, where all the newly added features are implemented. The major difference between the online and in-lab experiments is apparently the setting, in which participants are supposed to complete the experiment. Unlike the in-lab experiment, the online experiment allows participants to do the experiment wherever they would prefer to be. However, this factor is not an

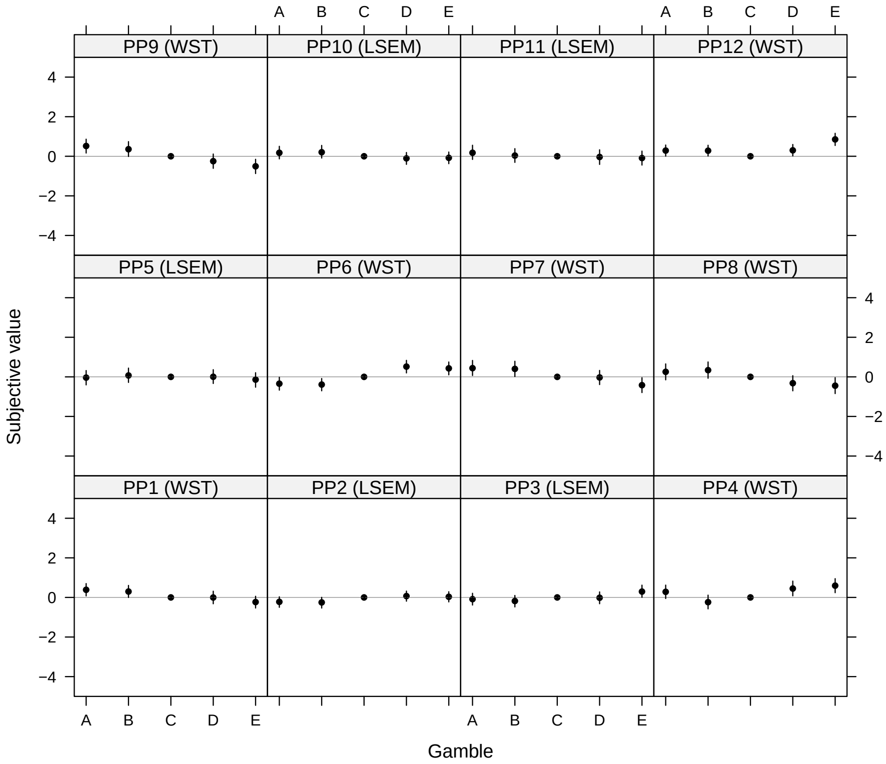


Figure 5.13: Subjective values of Gambles in Set 2 in in-lab data. Filled dots indicate posterior mean, and vertical lines superimposed on the dots indicate the 95% credible intervals.

advantage here, given that the current experiment is not particularly interesting, but long and tedious. Thus, the first thing I did after collecting the data was examine the collected RT to look for any contaminants in the data. Figure 5.15 shows the RT distribution of all the data, with participants, gamble pairs, and classifications all collapsed.

As shown in Figure 5.15, participants of the online experiment are much faster than those of other experiments. There are even a significant number of response time data faster than 0.3 second, which we would usually consider outliers. Indeed, as I examine the response time data from the online experiment, I encounter response times faster than 0.01 second quite frequently, which is technically impossible in practice. Even if a participant chooses a random gamble as soon as a trial begins, even without looking at it, it would surely take more than 0.01 second. While I was

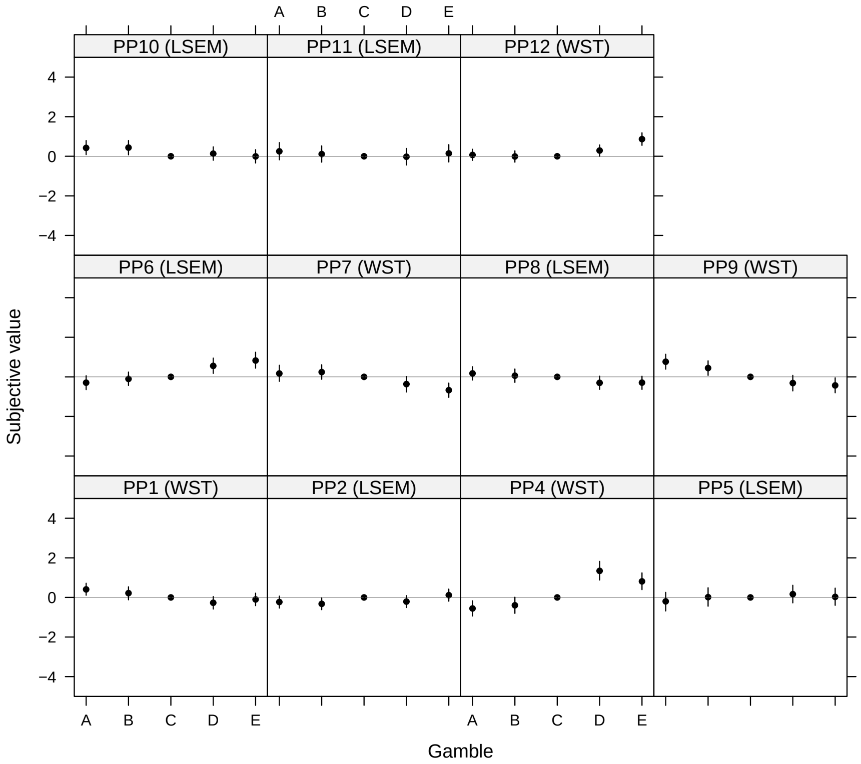


Figure 5.14: Subjective values of Gambles in Set 3 in in-lab data. Filled dots indicate posterior mean, and vertical lines superimposed on the dots indicate the 95% credible intervals.

finding out how such fast responses could be recorded, I realized that if one of the keys that I designate to choose gambles in this experiment is held down, a trial would end as soon as it begins and response time for that trial would be recorded faster than 0.01 second. So, non-cooperative participants would hold the key down for the whole experiment, because this is the fastest way they could complete the experiment.

Therefore, I apply the cutoff of 0.3 to the response time data to eliminate too fast responses from the analysis. And as a result, a great chunk of data, i.e., 10,102 response time data, are eliminated, where 3,368 response time data from Set 1, 3,352 response time data from Set 2, and 3,382 response time data from Set 3 are eliminated. This is a surprisingly big number of data. Given that one participant can yield 360 data points, 10,102 data are about 28 participant-worth data, which is even greater than the number of participants in the in-lab data. After excluding those data, I

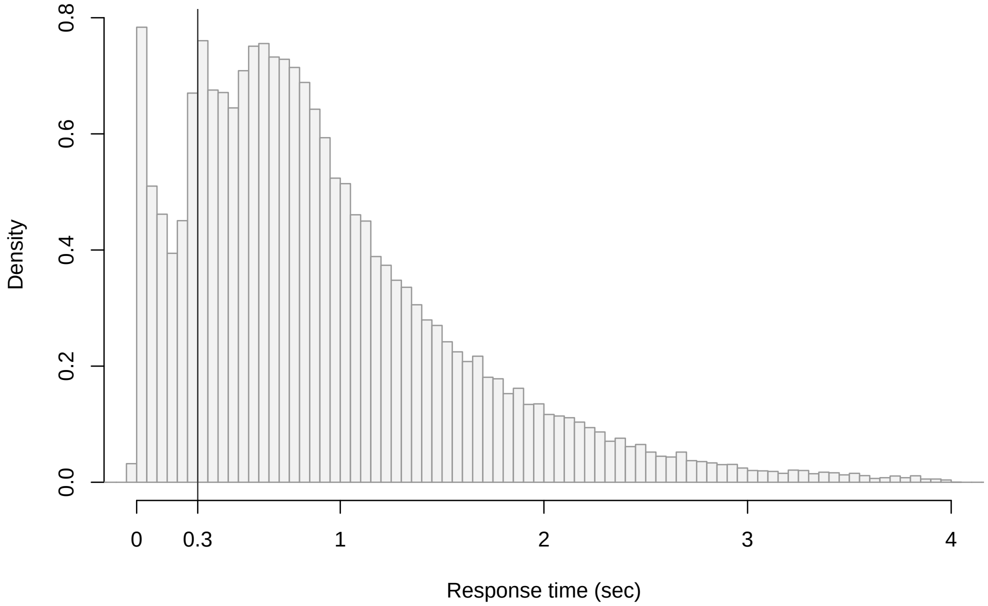


Figure 5.15: Histogram of observed RT from the online experiment. The vertical line indicates a cut-off value of 0.3 we adapted for other data sets.

reexamine the leftover data and decide to let go of the participants who lost more than 10% of the data as outliers. So the final data set looks as follows: the original number of participants was 173 for each gamble set, but after the elimination, now the number of participants are 81 for Gamble Set 1, 82 for Gamble Set 2, and 80 for Gamble Set 3. I proceed to the further analysis with these participants only.

5.3.2 Results

Classification analysis

As before, I classify participants into whether WST or LSEM, using the choice data only. Again, I compute the Bayes factors to classify the participants, by applying Kass and Raftery's (1995) decision criteria. The results are shown in Table 5.12. Due to the large number of participants, I choose not to list all the individual Bayes

Table 5.12: Summary of classifications for online data

	Number of participants			
Classification	Gamble Set 1	Gamble Set 2	Gamble Set 3	Total
WST	37	36	36	109
LSEM	44	46	44	134
NA	0	0	0	0
Total	81	82	80	243

factors for this data set, but only the summary table that includes the number of participants for each classification.

The WST and LSEM again perfectly account for the observed choice data patterns. One thing worth mentioning about the resulting classifications is that we observe more data sets classified to LSEM than to WST. In other words, more participants tend to integrate less information to produce a decision in this data set. It makes sense considering that the current data are from the online experiment. That is, the experiment has been administered not in a controlled setting, but in a setting participants would prefer, so there are higher chances that participants are surrounded by distractions when they do the experiment. Such distractions will surely make them hard to concentrate on the experiment, which I'd argue might have resulted in more LSEM participants than WST.

Diffusion model analysis

Posterior predictive check The same diffusion model analyses have been applied to the online data. First, I do the posterior predictive check of the model using the posterior predictive samples. This is done in two ways as before: Bayesian p-values for choice data, and density plots of the observed and predicted RT for RT data. I start with the Bayesian p-values. I compute the Bayesian p-values using the posterior predictive samples for all participants. Due to the large number of participants, however, I choose not to list all the Bayesian p-values for the participants; only the number of the p-values smaller than 0.05, implying lack of fit of the model, will be reported. For Gamble Set 1, 5 participants have p-values smaller than 0.05; for Gamble Set 2, 8 participants have p-values smaller than 0.05; for Gamble Set 3, 45 participants have p-values smaller than 0.05. Interestingly, the current diffusion model provides an adequate fit for Gamble Sets 1 and 2 data, but when it comes to Gamble Set 3 data, the number of the participants that are not well accounted for by the model is 45, which is more than half of the participants. More investigation of data would be needed to explain this surprising number of lack of fit with respect to Gamble Set 3 data, but for now I move on to the next analysis.

For response time data, I check the density plots of the observed response time (plotted as a histogram) and predicted response time (plotted as a density curve). The more the two densities are overlapped with each other, the better the model accounts for the response time data. Figures 5.16, 5.17, and 5.18 show the density plots of Gamble Sets 1, 2, and 3, respectively. All the density plots indicate that the current diffusion model provides a decent fit for the observed RT data. Even for Gamble Set 3, where the model couldn't account for more than half of the participants with respect to the choice data, the density plot (Figure 5.18) shows a decent fit for the RT data.

Bayesian inferences on the main parameters In this analysis, I examine the posterior estimates of the main parameters of the diffusion model, i.e., decision criterion a, and subjective values $\boldsymbol{u}$, to investigate the different cognitive processes assumed by transitivity and lexicographic semiorder. As before, for decision criterion a, the difference in group mean of the decision criterion between WST and LSEM is primarily investigated, because such differences are particularly related to the main hypothesis of the current study. For subjective values $\boldsymbol{u}$, I examine the trend of subjective values over the given gambles to determine if one forms intransitive preferences, represented by a cyclic pattern of preferences. The results regarding a are shown in Table 5.8 in the last three rows. The results of all three gamble sets indicate that the data from the online experiment don't support the main hypothesis regarding a. The two groups, i.e., WST and LSEM, show no difference in a, where the WST-classified participants in this data set have particularly low values of a, compared to other data sets. For example, in Table 5.8, posterior mean of μ_a^{WST} for Online Set 3 is estimated to be 1.89, lower than the posterior mean of μ_a^{LSEM} for Gamble Set 2 in Cavagnaro and Davis-Stober's (2014) data, which is 1.94. This effectively means that the WST-classified participants in the online experiment have accumulated less amount of information than the LSEM-classified participants in Cavagnaro

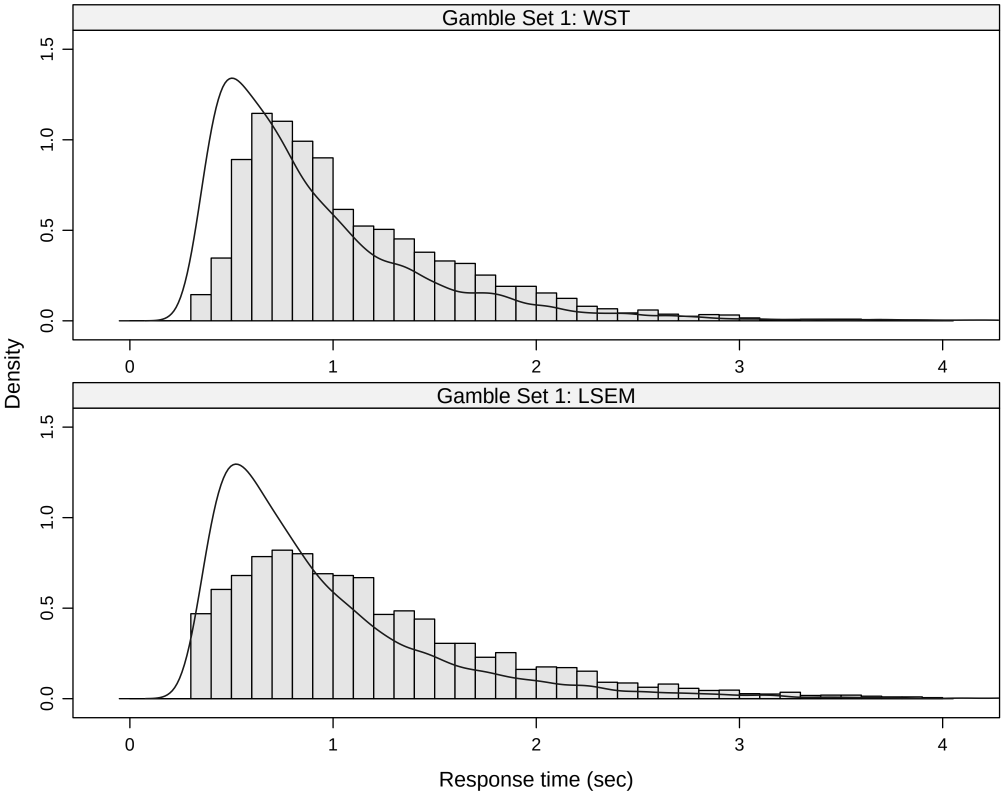


Figure 5.16: Density plots of RT for Gamble Set 1 in online data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve.

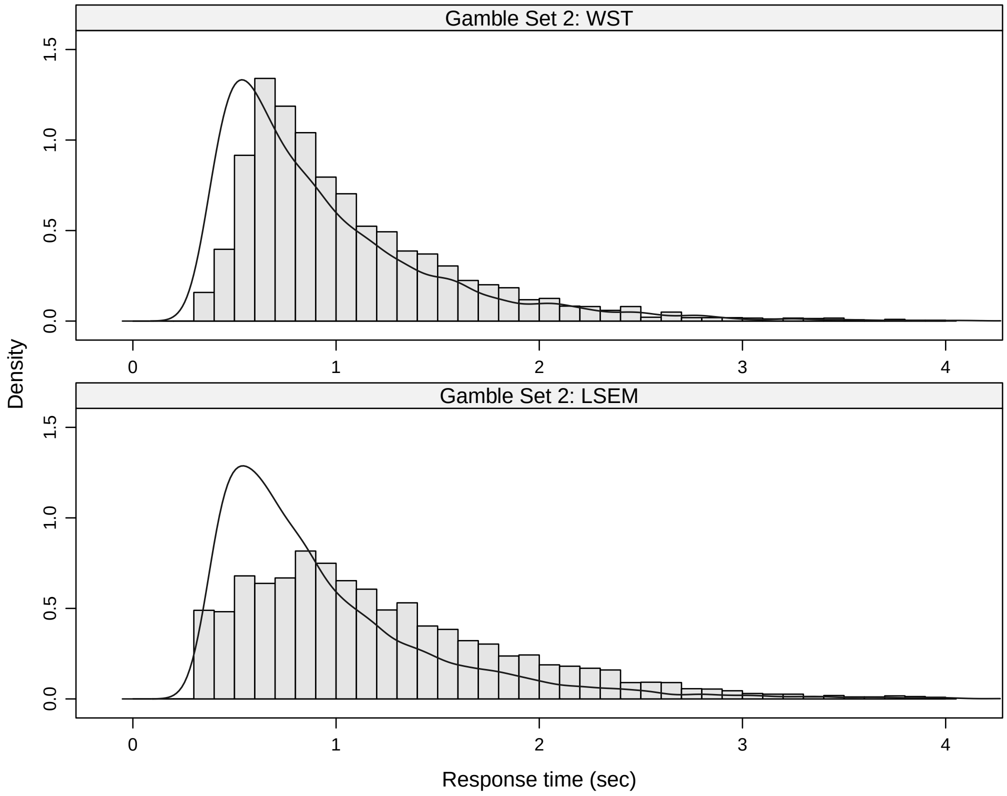


Figure 5.17: Density plots of RT for Gamble Set 2 in online data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve.

and Davis-Stober's experiment to produce a decision. As I mentioned before, this is likely due to the unique setting of the online experiment. Since the online experiment makes no requirement as to where participants are supposed to do the experiment, participants in the online experiment are likely to expose themselves to more distractions during the experiment. With distractions, participants are hardly focused on the experiment, which, in my opinion, results in such low values of μ_a^{WST} for all three gamble sets.

For subjective values of the gambles, as before, I examine the trend in which subjective values are estimated for the gambles. Due to the large number of participants, I choose not to include plots of subjective values. No interesting findings stand out in this data set; all LSEM subjective values have been estimated around 0, like other data sets, which implies cyclic patterns of preferences.

5.4 Discussion

The empirical analysis of the diffusion model demonstrated in this chapter that transitivity and lexicographic semiorder likely go through different cognitive processes, but such results depended on the setting where the data had been collected. First, for the data from Cavagnaro and Daivs-Stober's (2014) experiment, only Gamble Set 2 showed a significant difference in the decision criterion parameter a of the diffusion model between transitivity and lexicographic semiorder, yet Gamble Set 1 showed no significant difference in a between the two groups. I suspected that it's the way the participants were compensated in that experiment that affected the results, so I ended up designing my own experiment in order to address the problem, where participants now get to play the chosen gamble every trial. In the new experiment,

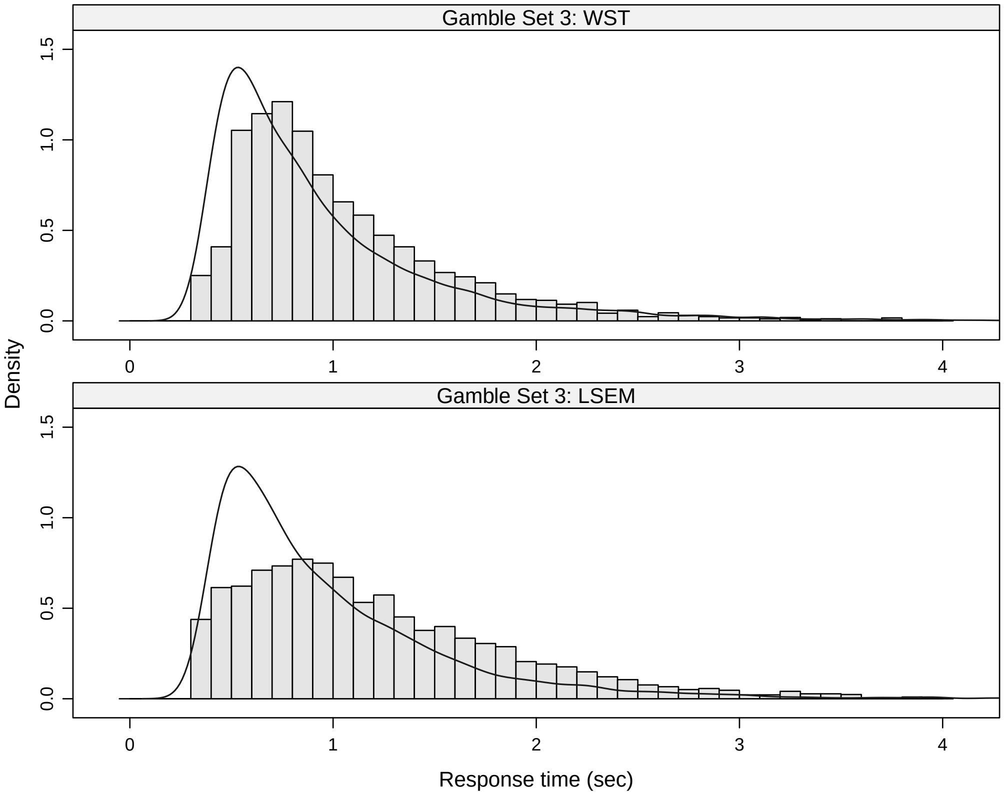


Figure 5.18: Density plots of RT for Gamble Set 3 in online data. Observed RTs are plotted as a histogram, predicted RTs are plotted as a density curve.

all the results from the diffusion model analysis were consistent with the hypothesis, implying that the two theories, transitivity and lexicographic semiorder, work based on two different cognitive processes; that is, those classified to transitivity tend to integrate more information than those classified to lexicographic semiorder to produce a decision. However, such results were not replicated when the same experiment was administered online. In the online experiment, participants were able to choose where they would do the experiment, so I suspect that more distractions must have been present when they did the online experiment than the in-lab experiment, which may hinder the transitivity participants from integrating more information.

As for the classification analysis, I employed the Weak Stochastic Transitivity (WST) and Lexicographic Semiorder Error Model (LSEM) as statistical models for transitivity and lexicographic semiorders, respectively. These two models were capable of accounting for almost all the observed choice patterns across all data sets, with only a few participants left unclassified. This is a remarkable result, in my opinion, given that the WST and LESM can only account for a small number of preference states, compared to the total number of preference states one is allowed to have. In the next chapter, I discuss the major implications of the current analyses, and how it has helped me answer the research questions that motivated me to conduct the whole study in the first place.

Chapter 6

General discussion

In this dissertation, I applied the diffusion model to preferential choice data to examine the underlying cognitive processes of transitivity and lexicographic semiorders. Transitivity is often considered a quality ascribe to a rational decision maker, while lexicographic semiorders are considered an alternative way to make decisions, especially when the given task is too daunting to apply transitive decision rules to. There are a number of factors that make a task difficult to deal with, including the number of pieces of information to process, and the need of making tradeoffs between conflicting attributes. If those factors make a task complicated, lexicographic semiorders can be immensely helpful here, because lexicographic semiorders don't require a decision maker to make tradeoffs between conflicting attributes, or even integrate all pieces of information to make a decision. Transitive decision rules, however, generally need to integrate all relevant pieces of information to make a decision because there is a possibility that the missing-out information could give rise to intransitive preferences. These two starkly different cognitive processes assumed by transitivity and lexicographic semiorders have rarely been compared to each other, empirically. Therefore, I am motivated to conduct the current study, where I aim to provide empirical evidence of two difference cognitive processes between transitivity and lexicographic semiorders, via the lens of the diffusion model.

In Chapter 1, I've reviewed related literature of the theories of transitivity and lexicographic semiorders. In this chapter, I specifically focused on how to test transitivity and lexicographic semiorder against empirical data, because those theories only make deterministic statements, while the data are variable (Tversky, 1969). Testing deterministic decision theories against the data has been considered a challenge as Luce and Narens (1994) put it. It requires one to build statistical models for the theory of interest so that the theory can account for variable choices, and also to develop an appropriate statistical testing method to test those statistical models against data. For statistical models, error model and mixture model have been widely considered, for statistical method for testing those models, the chi-bar-square statistics (Davis-Stober, 2009), Bayesian p-values (Myung et al., 2005), and Bayes factor (Kass & Raftery, 1995) have been widely considered.

In Chapter 2, I've reviewed the literature of the diffusion model (Ratcliff, 1978). The diffusion model is one of the most, if not the most, influential and widely applied cognitive models in cognitive psychology. The diffusion model has four main parameters: decision criterion, drift rate, starting point, and non-decision time. Each parameter in the diffusion model is particularly related to the respective underlying cognitive process, hence the diffusion model provides a straightforward way to investigate the underlying process via its parameters. However, the diffusion model had been considered only for perceptual choice tasks, not for preferential choice tasks. This is because preferential choice tasks had been thought to require different cognitive processes than the ones assumed by the diffusion model. But, recent neurophysiological findings suggest that even for preferential choice tasks, our brain would process the information in the same way as for perceptual choice tasks (Basten et al., 2010; Gold & Shadlen, 2007; Polanía et al., 2014). Such findings greatly inspired the current study, where I applied the diffusion model to preferential choice tasks. To this end, I reinterpreted the drift rates of the model and reparameterized it in terms of subjective values of the given alternatives. This specific approach I took to reparameterize drift rates has also been considered in other studies as well (e.g. Fudenberg et al., 2018; Konovalov & Krajbich, 2019; Webb et al., 2019).

In Chapter 3, I conducted simulation studies, particularly aimed at testing the given model's ability to recover the true data-generating parameter values. Note that the diffusion model I've modified for preferential choice data assumed no across-trial variability for any of the main parameters and fixed its starting point at 0.5, a point equidistant from both decision criteria. Since I placed constraints over the parameters of the diffusion model, I refer this diffusion model to the constrained diffusion model. In the simulation study, I generated two data sets, one from a diffusion model with no constraints over its parameters, and one from the constrained diffusion model I just described above. Then, I fitted the constrained diffusion model to both data sets and examined how well the given model could recover the true data-generating parameter values. The results suggest that the constrained diffusion model is able to recover the true data-generating parameter values to a great precision when the data were simulated from the constrained diffusion model. Even for the other data set, the constrained diffusion model showed a decent performance in terms of parameter recovery. Both results indicate that the given model is well capable of testing the main hypotheses of the current study.

In Chapter 4, I apply the given diffusion model to three real data sets, one from Cavagnaro and Davis-Stober's (2014) experiment, and two from my own experiments. Since the main purpose of the current study is to examine the underlying cognitive processes of transitivity and lexicographic semiorders, I first classify participants into either transitivity or lexicographic semiorders by computing Bayes factors. After the classification was done, the resulting classifications entered the diffusion model as a group variable, so that I could examine the difference in the main parameters between different classifications. The results suggest that participants classified to transitivity in general have a higher value of decision criterion a than those classified to lexicographic semiorders. The results varied depending on the setting the data were collected, which I discussed in detail in Chapter 4.

6.1 Back to research questions

The current study all started from one interesting observation: different people make different choices even for the same task. It sounds natural, but when we take this observation a little bit further, a set of interesting questions arise: How do those choices differ? and why do those choice differ? The present dissertation is like a journey for me, where I explore various decision theories to answer those two questions. During this journey, I chose transitivity and lexicographic semiorder theories to answer the how question, and the diffusion model to answer the why question. Below, I elaborate how those theories have helped me answer the research questions.

6.1.1 How do choices differ?

Different people make difference choices. This seems quite intuitive, because every people thinks differently. But, if we ask how those choices differ, it is not an easy question to answer, because it involves the choice of a model that can differentiate one pattern of observed choices from another. Also, we need to consider how we apply the chosen model to the observed data, because most decision theories predict deterministic relations, while the observed choice data can be variable. Hence, a successful model is the one that can account for observed choice variability, and at the same time, differentiate a pattern of choices from another. To this end, I chose transitivity and lexicographic semiorder. The two theories are unique in that they directly operate on the level of preferences. They don't assume any parameters that are related to a hypothetical decision-making process, like other economic theories, such as the expected utility theory (von Neumann & Morgenstern, 1947), or the cumulative prospect theory (Kahneman & Tversky, 1979). They rather make a direct statement about preferences, so I'd argue that they operate in a more fundamental way. For statistical models, I chose the Weak Stochastic Transitivity (WST) and the Lexicographic Semiorder Error Model (LSEM). Those are error models for transitivity and lexicographic semiorders, respectively, so those models assume that there exist a true preference behind observed choices, and that observed choice variability is due to an error.

As we've seen in Chapter 4, the WST and LSEM did an astounding job in classifying participants. Across all three data sets considered in Chapter 4, almost all participants were classified to either WST or LSEM, with just a few people left unclassified. Note that the WST and LSEM can only predict a small number of preferences in comparison to the total number of preferences one is allowed to have. With five alternatives, there are 1024 different preferences available for each individual. Of which, the WST can predict only 120 different preferences and the LSEM can predict 21 different preferences. In spite of such small numbers of preferences each model predicts, the WST and LSEM were able to account for almost all different choice patterns observed in the three data sets in Chapter 4. Such results imply the following points: first, different people make different choices not in a haphazard manner, but in a more systematic manner; second, such different choice patterns can be captured by either transitivity and lexicographic semiorders, that is, people would make choices with a transitive preference in mind, or they follow the lexicographic semiorder to make a decision.

6.1.2 Why do choices differ?

To answer why different people make different choices, I apply the diffusion model to choice and RT data. Being one of the cognitive models, the diffusion model offers a unique advantage in accounting for the underlying cognitive processes: the parameters of the diffusion model are directly related to various components of our cognitive processes. Hence, to learn what cognitive processes have brought about the observed choices, we simply examine its parameters, and this is what I've done in Chapter 4. First, I did the classification analysis to classify participants to either transitivity or lexicographic semiorders. These are the two models I chose to differentiate one choice pattern from another. Then, the resulting classifications enter the diffusion model as a group variable. In other words, I am interested in learning the underlying cognitive processes behind these different choice patterns.

As discussed in Chapter 4, the major difference between transitivity and lexicographic semiorders manifested via the decision criterion parameter a. That is, the participants classified to transitivity tended to accumulate more information to make decisions than those classified to lexicographic semiorders. Consider this result in the context of the second research question, why those choices differ. We first classified observed choices to transitivity or lexicographic semiorders. From this classification, we learn that there are at least two types of different choice patterns, beyond choice variability. And now the diffusion model analysis tells that such different choice patterns are deeply related to the amount of information needed to make a choice. In other words, one's willingness to integrate more information in making a decision is associated with which type of choice patterns he or she will end up with.

So, why do choices differ? Because different people have different levels of willingness to process more information for that choice. This is the answer I found from the perspective of the diffusion model analysis.

6.2 Limitations and future research

The current diffusion model has placed a number of constraints over its parameters to facilitate its estimation process. As seen in Chapter 3, however, this constrained model can yield misleading results when the data have more complexity than the model assumes. In this case, the diffusion model with less constraints over its parameters, such as the standard Ratcliff diffusion model (Ratcliff, 1978), might offer a better choice. Another limitation comes from the choice of the diffusion model itself. As we see in Chapter 4, the observed RT distribution of the LSEM-classified participants in the in-lab data seemed to have two distinct cognitive processes going on when they did the experiment. The diffusion model is not able to account for such bimodality of the distribution, and I think this is where I need to consider different cognitive models.

The future research is related to the point I just made as limitations. First, I consider the diffusion model with less constraints over the parameters for future research. In the current analysis, I frequently found a number of participants who showed lack of fit of the current model in terms of either choice or RT data. I suspect that the constraints I placed over the current model might be related to such lack of fit of the model. Second, I consider different cognitive models to examine the underlying cognitive processes of transitivity and lexicographic semiorders. As of now, there are a good number of cognitive models out there, including the leaky competing accumulator model (Usher & McClelland, 2001), Ornstein-Uhlenbeck process model, or the shifted Wald model (Anders et al., 2016). Those models have their own pros and cons, but depending on the type of data one of these models can provide a better fit than the diffusion model.

Appendix A

Stan code for the diffusion model analysis

```
functions {
  real wiener_diffusion_lpdf(real y, int dec, real alpha,
                             real tau, real beta, real delta) {
    /* Wiener diffusion density parameters
     * y: response time
     * dec: corresponding choice made (1: payoff-favored gamble,
     *                                 0: probability-favored gamble)
     * alpha: boundary separation
     * tau: non-decision time
     * beta: bias
     * delta: drift rate
     */
    if (dec == 1) {
      return wiener_lpdf(y | alpha, tau, beta, delta);
    } else {
      return wiener_lpdf(y | alpha, tau, 1 - beta, - delta);
    }
  }

  matrix cov_GPL2(matrix x, real sq_eta, real sq_rho, real delta) {
    int N = dims(x)[1];
    matrix[N, N] K;
    for (i in 1:(N-1)) {
      K[i, i] = sq_eta + delta;
      for (j in (i + 1):N) {
        K[i, j] = sq_eta * exp(-sq_rho * square(x[i,j]) );
        K[j, i] = K[i, j];
      }
    }
    K[N, N] = sq_eta + delta;
    return K;
  }
}

data {
  int<lower=1> P;                // Number of subjects
  int<lower=1> J;                // Total number of observations
  int<lower=1> K;                // Number of unique gambles (K = 5)
  int<lower=1> G;                // Number of groups
  int<lower=1> pp[J];            // Subject on Observation n
  int<lower=1, upper=2> gg[P];   // Resulting group by classification
  vector[P] X_tr;                // Resulting transitivity classification on observation n
  row_vector[K] X_u[J];          // Design matrix for v
  matrix[K, K] dist_mat;
  real const_z;                  // Bias parameter set to .5 (i.e., unbiased decision)
  real y[J];                     // Response time
  int<lower=0, upper=1> dec[J]; // Choice made in each trial
  real min_rt[P];                // Minimum response time for each subject
}

parameters {
  real<lower=0> beta0_a;
  real beta1_a;
  vector[P] stn_a;
  real<lower=0> sigma_a; // prior scale for alpha
  vector[K] u[P];
  real mu_t;
  real<lower=0> sigma_t;
  vector[P] stn_t;
  vector<lower=0>[G] etasq;
  vector<lower=0>[G] rhosq;
  vector<lower=0>[G] sigmasq;
}
```

```
transformed parameters {
  vector<lower=0>[G] mu_a;
  vector[P] a;               // Boundary separation parameter
  vector[J] v;               // Drift rate
  matrix[K, K] Sigma[G];
  vector[P] t_er;                 // Non-decision time

  // Boundary separation parameter a
  for (j in 1:J) {
    mu_a[gg[pp[j]]] = beta0_a + beta1_a*X_tr[pp[j]];
    a[pp[j]] = mu_a[gg[pp[j]]] + stn_a[pp[j]]*sigma_a;
    a[pp[j]] = exp(a[pp[j]]);   // log-transformation of a
  }

  // Drift rate v
  for (j in 1:J) {
    v[j] = X_u[j] * u[pp[j]];
  }
  for (g in 1:G) {
    Sigma[g] = cov_GPL2(dist_mat, etasq[g], rhosq[g], sigmasq[g]);
  }

  // non-decision time t_er
  for (p in 1:P){
    t_er[p] = Phi_approx(mu_t + stn_t[p]*sigma_t) * min_rt[p];
  }
}

model {
  beta0_a ~ normal(0.5, 0.5);
  beta1_a ~ normal(0, 0.5);
  sigma_a ~ exponential(0.5);
  stn_a ~ normal(0, 1);

  etasq ~ exponential(2);
  rhosq ~ exponential(1);
  sigmasq ~ exponential(0.5);
  for (p in 1:P) {
    u[p] ~ multi_normal(rep_vector(0, K), Sigma[gg[p]]);
  }
  mu_t ~ normal(0, 1);
  sigma_t ~ exponential(1);
  stn_t ~ normal(0, 1);

  for (j in 1:J){
    y[j] ~ wiener_diffusion(dec[j], a[pp[j]], t_er[pp[j]], const_z, v[j]);
  }
}
```

References

Amit, D. J., & Brunel, N. (1997). Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex*, *7*(3), 237–252. doi: 10.1093/cercor/7.3.237

Anand, P. (1993). The Philosophy of Intransitive Preference. *The Economic Journal*, *103*(417), 337–346. doi: 10.2307/2234772

Anders, R., Alario, F.-X., & Van Maanen, L. (2016). The shifted Wald distribution for response time data analysis. *Psychological Methods*, *21*(3), 309–327. doi: 10.1037/met0000066

Annis, J., Miller, B. J., & Palmeri, T. J. (2017). Bayesian inference with Stan: A tutorial on adding custom distributions. *Behavior Research Methods*, *49*(3), 863–886. doi: 10.3758/s13428-016-0746-9

Armstrong, W. E. (1950). A note on the theory of consumer's behavior. *Oxford Economic Papers*, *2*, 119–122.

Aschenbrenner, A. J., Balota, D. A., Gordon, B. A., Ratcliff, R., & Morris, J. C. (2016). A diffusion model analysis of episodic recognition in preclinical individuals with a family history for Alzheimer's disease: The adult children study. *Neuropsychology*, *30*(2), 225–238. doi: 10.1037/neu0000222

Bar-Hillel, M., & Margalit, A. (1988). How vicious are cycles of intransitive choice? *Theory and Decision. An International Journal for Multidisciplinary Advances in Decision Science*, *24*(2), 119–145. doi: 10.1007/BF00132458

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*, 1281–1311.

Basten, U., Biele, G., Heekeren, H. R., & Fiebach, C. J. (2010). How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences*, *107*(50), 21767–21772. doi: 10.1073/pnas.0908104107

Becker, G. M., Degroot, M. H., & Marschak, J. (1963). Stochastic models of choice behavior. *Behavioral Science*, *8*(1), 41–55. doi: 10.1002/bs.3830080106

Bentham, J. (1789). *The principles of morals and legislation.* London.

Betancourt, M. (2016). *Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo.* doi: 10.48550/arXiv.1604.00695

Birnbaum, M. H., & Gutierrez, R. J. (2007). Testing for intransitivity of preferences predicted by a lexicographic semi-order. *Organizational Behavior and Human Decision Processes*, *104*, 96–112.

Block, H. D., & Marschak, J. (1960). Random orderings and stochastic theories of responses. In I. Olkin, S. Ghurye, W. Hoefding, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 97–132). Stanford, CA: Stanford University Press.

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., ... Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46–75. doi: 10.1016/j.jmp.2018.09.004

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. L. (2007). Extending a biologically inspired model of choice: Multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1655–1670. doi: 10.1098/rstb.2007.2059

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed–accuracy tradeoff. *Trends in Neurosciences*, *33*(1), 10–16. doi: 10.1016/j.tins.2009.09.002

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, *113*(2), 409–432. doi: 10.1037/0033-295X.113.2.409

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. doi: 10.1016/j.cogpsych.2007.12.002

Camerer, C. (2013). Goals, methods and progress in neuroeconomics. *Annual Review of Economics*, *5*, 425–455. doi: 10.1146/annurev-economics-082012-123040

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). *Stan* : A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). doi: 10.18637/jss.v076.i01

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, *39*(2), 83–87. doi: 10.1080/00031305.1985.10479400

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167–174.

Cavagnaro, D. R., & Davis-Stober, C. P. (2014). Transitive in our preferences, but transitive in different ways: An analysis of choice variability. *Decision*, *1*(2), 102–122. doi: 10.1037/dec0000011

Clithero, J. A. (2018). Improving out-of-sample predictions using response times and a model of the decision process. *Journal of Economic Behavior & Organization*, *148*, 344–375. doi: 10.1016/j.jebo.2018.02.007

Colonius, H., & Marley, A. A. (2015). Decision and Choice: Random Utility Models of Choice and Response Time. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 901–905). Elsevier. doi: 10.1016/B978-0-08-097086-8.43033-3

Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology*, *53*(1), 1–13. doi: 10.1016/j.jmp.2008.08.003

Davis-Stober, C. P. (2010). A bijection between a set of lexicographic semiorders and pairs of non-crossing Dyck paths. *Journal of Mathematical Psychology*, *54*(6), 471–474.

Davis-Stober, C. P. (2012). A lexicographic semiorder polytope and probabilistic representations of choice. *Journal of Mathematical Psychology*, *56*(2), 86–94.

Davis-Stober, C. P., Brown, N., & Cavagnaro, D. R. (2015). Individual differences in the algebraic structure of preferences. *Journal of Mathematical Psychology*, *66*, 70–82. doi: 10.1016/j.jmp.2014.12.003

Davis-Stober, C. P., McCarthy, D. M., Cavagnaro, D. R., Price, M., Brown, N., & Park, S. (2019). Is cognitive impairment related to violations of rationality? A laboratory alcohol intoxication study testing transitivity of preference. *Decision*, *6*(2), 134–144.

Davis-Stober, C. P., Park, S., Brown, N., & Regenwetter, M. (2016). Reported violations of rationality may be aggregation artifacts. *Proceedings of the National Academy of Sciences*, *113*(33), E4761-E4763. doi: 10.1073/pnas.1606997113

Donders, F. C. (1969). Over de snelheid van psychische processen (On the speed of psychological processes). In *Attention and Performance II* (W. Koster ed.). Amsterdam: North-Holland.

Donzallaz, M. C., Haaf, J. M., & Stevenson, C. E. (2022). Creative or not? Hierarchical diffusion modeling of the creative evaluation process. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *Advance online publication*. doi: 10.1037/xlm0001177

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics Letters B*, *195*, 216–222.

Dutilh, G., & Rieskamp, J. (2016). Comparing perceptual and preferential decision making. *Psychonomic Bulletin & Review*, *23*(3), 723–737. doi: 10.3758/s13423-015-0941-1

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. doi: 10.1073/pnas.1818430116

Fechner, G. (1860). *Element of Psychophysics*. New York: Holt, Rinehart and Winston.

Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed., Vol. 1). New York: Wiley.

Fishburn, P. C. (1974). Lexicographic orders, utilities and decision rules: A survey. *Management Science*, *20*(11), 1442–1471. doi: 10.1287/mnsc.20.11.1442

Fudenberg, D., Strack, P., & Strzalecki, T. (2018). Speed, Accuracy, and the Optimal Timing of Choices. *American Economic Review*, *108*(12), 3651–3684. doi: 10.1257/aer.20150742

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelfand, A. E., Smith, A. F. M., & Lee, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*(418), 523–532.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (Third ed.). New York: Chapman and Hall/CRC. doi: 10.1201/9780429258411

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, MA: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511.

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, *1*(1), 107–143. doi: 10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the Fast and Frugal Way: Models of Bounded Rationality. *Psychological Review*, *103*(4), 650–669.

Gigerenzer, G., & Selten, R. (Eds.). (2001). *Bounded rationality: The adaptive toolbox.* Cambridge, MA: MIT press.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart.* Oxford University Press.

Gill, D., & Prowse, V. L. (2017). Using Response Times to Measure Strategic Complexity and the Value of Thinking in Games. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2902411

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*(1), 10–16. doi: 10.1016/S1364-6613(00)01567-9

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*(1), 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Grasman, R. P., Wagenmakers, E.-J., & van der Maas, H. L. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology*, *53*(2), 55–68. doi: 10.1016/j.jmp.2009.01.006

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Harless, D. W., & Camerer, C. F. (1994). The Predictive Utility of Generalized

Expected Utility Theories. *Econometrica : journal of the Econometric Society*, *62*(6), 1251–1290. doi: 10.2307/2951749

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. doi: 10.1093/biomet/57.1.97

Heathcote, A., Popiel, S., & Mewhort, D. (1991). Analysis of response time distributions: An example using the Stroop Task. *Psychological Bulletin*, *109*, 340–347. doi: 10.1037/0033-2909.109.2.340

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology*, *91*, 70–87. doi: 10.1016/j.jmp.2019.03.004

Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience*, *9*(6), 467–479. doi: 10.1038/nrn2374

Hey, J. D. (2005). Why we should not be silent about noise. *Experimental Economics*, *8*, 325–345.

Hey, J. D., & Orme, C. (1994). Investigating Generalizations of Expected Utility Theory Using Experimental Data. *Econometrica : journal of the Econometric Society*, *62*(6), 1291. doi: 10.2307/2951750

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

Irwin, F. W. (1958). An analysis of the concepts of discrimination and preference. *The American Journal of Psychology*, *71*, 152–163.

Iverson, G., & Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, *10*, 131–153.

Jackman, S. (2000). Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo. *American Journal of Political Science*, *44*(2), 375. doi: 10.2307/2669318

Jackman, S. (2009). *Bayesian analysis for the social sciences.* John Wiley and Sons.

Jastrow, J. (1890). *The time relations of mental phenomena.* New York: Hodges.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica : journal of the Econometric Society*, *47*(2), 263. doi: 10.2307/1914185

Karabatsos, G., & Batchelder, W. H. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika*, *68*(3), 373–389.

Karabatsos, G., & Sheu, C.-F. (2004). Order-Constrained Bayes Inference for Dichotomous Models of Unidimensional Nonparametric IRT. *Applied Psychological Measurement*, *28*(2), 110–125. doi: 10.1177/0146621603260678

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 10.2307/2291091

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value trade-offs.* Cambridge University Press. doi: 10.1017/CBO9781139174084

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*(12), 6367–6379. doi: 10.1016/j.csda.2007.01.024

Konovalov, A., & Krajbich, I. (2019). Revealed strength of preference: Inference from response times. *Judgment and Decision Making*, *14*(4), 381–394. doi: 10.2139/ssrn.3024233

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., & Fehr, E. (2015). A common mechanism underlying food choice and social decisions. *PLOS Computational Biology*, *11*(10), e1004371. doi: 10.1371/journal.pcbi.1004371

LaBerge, D. A. (1962). A recruitment theory of simple behavior. *Psychometrika*, *27*, 375–396.

Laming, D. R. (1968). *Information theory of choice reaction time.* New York: Wiley.

Lerche, V., & Voss, A. (2016). Model Complexity in Diffusion Modeling: Benefits of Making the Model More Parsimonious. *Frontiers in Psychology*, *7*. doi: 10.3389/fpsyg.2016.01324

Lerche, V., & Voss, A. (2017). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, *81*(3), 629–652. doi: 10.1007/s00426-016-0770-5

Loomes, G. (2005). Modelling the Stochastic Component of Behaviour in Experiments: Some Issues for the Interpretation of Data. *Experimental Economics*, *8*(4), 301–323. doi: 10.1007/s10683-005-5372-9

Loomes, G., Moffatt, P. G., & Sugden, R. (2002). A Microeconometric Test of Alternative Stochastic Theories of Risky Choice. *Journal of Risk and Uncertainty*, *24*(2), 103–130.

Loomes, G., & Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, *39*(3-4), 641–648. doi: 10.1016/0014-2921(94)00071-7

Loomes, G., & Sugden, R. (1998). Testing Different Stochastic Specificationsof Risky Choice. *Economica*, *65*(260), 581–598. doi: 10.1111/1468-0335.00147

Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica : journal of the Econometric Society*, *24*(2), 178–191.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Luce, R. D., & Narens, L. (1994). Fifteen Problems Concerning the Representational Theory of Measurement. In P. Humphreys (Ed.), *Patrick Suppes: Scientific Philosopher* (pp. 219–249). Dordrecht: Springer Netherlands.

Luce, R. D., & Raiffa, H. (1989). *Games and decisions: Introduction and critical* *survey*. Courier Corporation.

Luce, R. D., & Suppes, P. (1965). Preference, Utility, and Subjective Probability. In *Handbook of mathematical psychology* (Vol. 3). New York: Wiley.

Marley, A. A., & Colonius, H. (1992). The "Horse Race" random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology*, *36*, 1–20.

Masatlioglu, Y., Nakajima, D., & Ozbay, E. Y. (2012). Revealed Attention. *American Economic Review*, *102*(5), 2183–2205. doi: 10.1257/aer.102.5.2183

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*(5), 798–817. doi: 10.3758/PBR.16.5.798

May, K. O. (1954). Intransitivity, Utility, and the Aggregation of Preference Patterns. *Econometrica : journal of the Econometric Society*, *22*(1), 1. doi: 10.2307/1909827

McClelland, G. (1978). *Equal versus differential weighting for multiattribute decisions: There are no free lunches* (Center Report No. 207). University of Colorado, Boulder, CO: Institute of Cognitive Science.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models* (Second ed.). London: Chapman and Hall/CRC. doi: 10.1201/9780203753736

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (Second ed.). CRC press.

Mcfadden, D. (2001). Economic Choices. *The American Economic Review*, *91*(3), 351–378.

Mellers, B. A., & Biagini, K. (1994). Similarity and choice. *Psychological Review*, *101*(3), 505–518.

Meng, X.-L. (1994). Posterior Predictive p-Values. *The Annals of Statistics*, *22*(3). doi: 10.1214/aos/1176325622

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. doi: 10.1063/1.1699114

Miller, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 539–543. doi: 10.1037/0096-1523.14.3.539

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making*, *5*(6), 437–449. doi: 10.1017/S1930297500001285

Moustafa, A. A., Kéri, S., Somlai, Z., Balsdon, T., Frydecka, D., Misiak, B., & White, C. (2015). Drift diffusion model of reward and punishment learning in schizophrenia: Modeling and experimental data. *Behavioural Brain Research*, *291*, 147–154. doi: 10.1016/j.bbr.2015.05.024

Mulder, M. J., Bos, D., Weusten, J. M., van Belle, J., van Dijk, S. C., Simen, P., … Durston, S. (2010). Basic impairments in regulating the speed-accuracy tradeoff predict symptoms of attention-deficit/hyperactivity disorder. *Biological Psychiatry*, *68*(12), 1114–1119. doi: 10.1016/j.biopsych.2010.07.031

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. doi: 10.1016/S0022-2496(02)00028-7

Myung, I. J., Karabatsos, G., & Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, *49*(3), 205–225. doi: 10.1016/j.jmp.2005.02.004

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95.

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, *53*(4), 222–230. doi: 10.1016/j.jmp.2009.02.003

Neal, R. M. (2011). MCMC using hamiltonian dynamics. In S. Brooks, A. Gelman, G. L. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo.* Chapman and Hall/CRC.

Pachella, R. G. (1974). The interpretation of reaction time in information-processing research. In *Human information processing: Tutorials in performance and cognition.* New Jersey: Lawrence Erlbaum Associates Publishers. doi: 10.4324/9781003176688-2

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, *5*(5), 376–404. doi: 10.1167/5.5.1

Park, S., Davis-Stober, C. P., Snyder, H. K., Messner, W., & Regenwetter, M. (2019). Cognitive Aging and Tests of Rationality. *The Spanish Journal of Psychology*, *22*, E57. doi: 10.1017/sjp.2019.52

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology*, *43*, 87–131.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge University Press.

Polanía, R., Krajbich, I., Grueschow, M., & Ruff, C. C. (2014). Neural oscillations and synchronization differentially support evidence accumulation in Perceptual and value-based decision making. *Neuron*, *82*(3), 709–720. doi: 10.1016/j.neuron.2014.03.014

R Core Team. (2022). *R: A language and environment for statistical computing* [Manual]. Vienna, Austria.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. doi: 10.1037/0033-295X.85.2.59

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*(2), 278–291. doi: 10.3758/BF03196283

Ratcliff, R. (2008). The EZ diffusion method: Too EZ? *Psychonomic Bulletin & Review*, *15*(6), 1218–1228. doi: 10.3758/PBR.15.6.1218

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*(1), 159–182. doi: 10.1037/0033-295X.111.1.159

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922. doi: 10.1162/neco.2008.12-06-420

Ratcliff, R., Perea, M., Colangelo, A., & Buchanan, L. (2004). A diffusion model account of normal and impaired readers. *Brain and Cognition*, *55*(2), 374–382. doi: 10.1016/j.bandc.2004.02.051

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science*, *9*(5), 347–356. doi: 10.1111/1467-9280.00067

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(1), 127–140.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*(2), 333–367. doi: 10.1037/0033-295X.111.2.333

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. doi: 10.1016/j.tics.2016.01.007

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging*, *19*(2), 278–289. doi: 10.1037/0882-7974.19.2.278

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging*, *16*(2), 323–341. doi: 10.1037/0882-7974.16.2.323

Ratcliff, R., Thapar, A., & Mckoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics*, *65*(4), 523–535. doi: 10.3758/BF03194580

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50*(4), 408–424. doi: 10.1016/j.jml.2003.11.002

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*(3), 127–157. doi: 10.1016/j.cogpsych.2009.09.001

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481. doi: 10.3758/BF03196302

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*(2), 261–300. doi: 10.1037/0033-295X.106.2.261

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2010). Testing Transitivity of Preferences on Two-Alternative Forced Choice Data. *Frontiers in Psychology*, *1*. doi: 10.3389/fpsyg.2010.00148

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of Preferences. *Psychological Review*, *118*(1), 42–56.

Regenwetter, M., Davis-Stober, C. P., Lim, S. H., Guo, Y., Popova, A., Zwilling, C., … Messner, W. (2014). QTest: Quantitative testing of theories of binary choice. *Decision*, *1*(1), 2–34. doi: 10.1037/dec0000007

Resnik, M. D. (1987). *Choices: An introduction to decision theory.* University of Minnesota Press.

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, *127*(2), 258–276. doi: 10.1016/j.actpsy.2007.05.004

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223. doi: 10.3758/BF03257252

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The Lognormal Race: A Cognitive-Process Model of Choice and Latency with Desirable Psychometric Properties. *Psychometrika*, *80*(2), 491–513. doi: 10.1007/s11336-013-9396-3

Rousselet, G. A., & Wilcox, R. R. (2020). Reaction Times and other Skewed Distributions. *Meta-Psychology*, *4*. doi: 10.15626/MP.2019.1630

Savage, L. J. (1954). *The foundations of statistics.* New York: John Wiley and Sons.

Schall, J. D. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience*, *2*(1), 33–42. doi: 10.1038/35049054

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*(3), 414–429. doi: 10.1037/0096-3445.136.3.414

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers*, *33*(4), 457–469.

Shadlen, M. N., & Shohamy, D. (2016). Decision Making and Sequential Sampling from Memory. *Neuron*, *90*(5), 927–939. doi: 10.1016/j.neuron.2016.04.036

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*(2), 207–222. doi: 10.1037/0033-2909.134.2.207

Shanteau, J., & Thomas, R. P. (2000). Fast and frugal heuristics: What about unfriendly enviornments? *Behavioral and Brain Sciences*, *23*(5), 762–763. doi: 10.1017/S0140525X00003447

Shepard, R. N. (1964). On subjectively optimum selections among multi-attribute alternatives. In *In Human Judgments and Optimality* (M. W. Shelby, G. L. Bryan ed., pp. 257–281). New York: Wiley.

Shepsle, K. A., & Bonchek, M. S. (1997). *Analyzing politics: Rationality, behavior, and institutions.* New York: Norton.

Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: Inequality, order, and shape restrictions.* New York: John Wiley and Sons.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. doi: 10.2307/1884852

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129–138.

Simon, H. A. (1978). Rationality as process and as product of thought. *American Economic Review*, *68*(2), 1–16.

Simon, H. A. (1990). Invariants of Human Behavior. *Annual Review of Psychology*, *41*(1), 1–20. doi: 10.1146/annurev.ps.41.020190.000245

Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review*, *116*(2), 283–317. doi: 10.1037/a0015156

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. doi: 10.1111/1467-9868.00353

Spiliopoulos, L., & Ortmann, A. (2018). The BCD of response time analysis in experimental economics. *Experimental Economics*, *21*(2), 383–433. doi: 10.1007/s10683-017-9528-1

Stan Development Team. (2022). *Stan modeling language users guide and reference manual, 2.31. https://mc-stan.org.*

Starmer, C. (2000). Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk. *Journal of Economic Literature*, *38*(2), 332–382.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*(3), 251–260.

Summerfield, C., & Tsetsos, K. (2012). Building Bridges between Perceptual and Economic Decision-Making: Neural and Computational Mechanisms. *Frontiers in Neuroscience*, *6*. doi: 10.3389/fnins.2012.00070

Taussig, F. (1912). *Principles of economics.* New York: Macmillan.

Todd, P. M., & Gigerenzer, G. (2000). Précis of *Simple Heuristics that make us smart*. *Behavioral and Brain Sciences*, *23*(5), 727–741. doi: 10.1017/S0140525X00003447

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes.* New York: Cambridge University Press.

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, *113*(11), 3102–3107. doi: 10.1073/pnas.1519157113

Tuerlinckx, F., Maris, E., Ratcliff, R., & De Boeck, P. (2001). A comparison of four methods for simulating the diffusion process. *Behavior Research Methods, Instruments, & Computers*, *33*(4), 443–456. doi: 10.3758/BF03195402

Tullock, G. (1964). The irrationality of intransitivity. *Oxford Economic Papers*, *16*, 401–406.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31–48. doi: 10.1037/h0026750

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, *95*(3), 371–384. doi: 10.1037/0033-295X.95.3.371

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592.

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2017). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review*, *24*(2), 547–556. doi: 10.3758/s13423-016-1081-y

van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology*, *53*(6), 463–473. doi: 10.1016/j.jmp.2009.09.004

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*(6), 1011–1026. doi: 10.3758/BF03193087

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, *16*(1), 44–62. doi: 10.1037/a0021765

Vandekerckhove, J., Verheyen, S., & Tuerlinckx, F. (2010). A crossed random effects diffusion model for speeded semantic categorization decisions. *Acta Psychologica*, *133*(3), 269–282. doi: 10.1016/j.actpsy.2009.10.009

von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior* (Second ed.). Princeton, NJ: Princeton University Press.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*(7), 1206–1220.

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*(4), 767–775. doi: 10.3758/BF03192967

Wabersich, D., & Vandekerckhove, J. (2014). The RWiener Package: An R Package Providing Distribution Functions for the Wiener Diffusion Model. *The R Journal*, *6*(1), 49–56. doi: 10.32614/RJ-2014-005

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi: 10.3758/BF03194105

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140–159. doi: 10.1016/j.jml.2007.04.006

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review*, *15*(6), 1229–1235. doi: 10.3758/PBR.15.6.1229

Wagenmakers, E.-J., Van Der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22. doi: 10.3758/BF03194023

Wang, X.-J. (2002). Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. *Neuron*, *36*(5), 955–968. doi: 10.1016/S0896-6273(02)01092-9

Webb, R., Levy, I., Lazzaro, S. C., Rutledge, R. B., & Glimcher, P. W. (2019). Neural random utility: Relating cardinal neural observables to stochastic choice behavior. *Journal of Neuroscience, Psychology, and Economics*, *12*(1), 45–72. doi: 10.1037/npe0000101

Weigard, A., & Huang-Pollock, C. (2014). A diffusion modeling approach to understanding contextual cueing effects in children with ADHD. *Journal of Child Psychology and Psychiatry*, *55*(12), 1336–1344. doi: 10.1111/jcpp.12250

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85. doi: 10.1016/0001-6918(77)90012-9

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*. doi: 10.3389/fninf.2013.00014

Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., & Rueckl, J. (2015). Responding to nonwords in the lexical decision task: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 597–613. doi: 10.1037/xlm0000064

Zeguers, M. H., Snellings, P., Tijms, J., Weeda, W. D., Tamboer, P., Bexkens, A., & Huizenga, H. M. (2011). Specifying theories of developmental dyslexia: A diffusion model analysis of word recognition. *Developmental Science*, *14*(6), 1340–1354. doi: 10.1111/j.1467-7687.2011.01091.x

Zwilling, C. E., Cavagnaro, D. R., Regenwetter, M., Lim, S. H., Fields, B., & Zhang, Y. (2019). QTest 2.1: Quantitative testing of theories of binary choice using Bayesian inference. *Journal of Mathematical Psychology*, *91*, 176–194. doi: 10.1016/j.jmp.2019.05.002

VITA

Sanghyuk Park was born and raised in Seoul, South Korea. Before attending the University of Missouri in Columbia for his doctoral degree, he attended Yonsei University, Seoul, South Korea, where he earned his bachelor's degree in Psychology and the degree of Master of Science in industrial and organizational psychology. While he was in his master's program in Yonsei University, he had an opportunity to get involved in a project, where the goal of the project was to develop an aptitude test for commercial drivers in South Korea. This was where he grew his interest in statistics and the decision-making in general. He decided to study abroad to pursue further his studies in the topics and ended up working with Dr. Davis-Stober at the University of Missouri in Columbia, where he earned his degree of Doctor of Philosophy in Quantitative Psychology. His research interest lies in computational modeling of the decision-making for risky choices. Specifically, he has particular interest in Bayesian statistics and sequential sampling models, which includes the diffusion model. Indeed, his dissertation topic is centered on applying the diffusion model analysis to decision-making models, using the Bayesian statistics.