Decision Making under Uncertainty: Revealing, Characterizing and Modeling Individual Differences in the Iowa Gambling Task.

by

Lee I Newman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering and Psychology)
in The University of Michigan
2009

Doctoral Committee:

Professor Thad A. Polk, Chair
Professor Satinder Singh Baveja
Professor John E. Laird
Professor Richard L. Lewis
Professor David E. Meyer

Ellen…
for her support, thoughtful feedback and patient and willing service as a human whiteboard.

My parents…
who have offered unwavering (but at times concerned!) support throughout the many twists in the scholarly and professional road I have trodden.

"Poppey" Rudolph…
for the tool locker that inspired, and the life lesson that happiness and simplicity are inseparable.

Jack "O" Orchard…
who demonstrated how horrific losses may become immeasurable gains when one refuses to give up hope.

ACKNOWLEDGMENTS

First and foremost, I want to thank Thad Polk who has been my advisor, research mentor, teaching mentor, and academic coach extraordinaire over the course of the last seven years. Returning to school after a decade away, to pursue a degree in two fields which I knew nothing about has been an incredibly challenging experience. Thad's perpetual cheeriness, reassuring style, and eternal willingness to offer help were often the fuel that kept me going, particularly in moments of doubt and uncertainty. It is said that influence through modeling trumps the more common tools of telling, consulting and logically persuading. And it has been demonstrated that mere exposure is often sufficient to shape behavior. Thad has certainly been an outstanding mentor in the active sense, but he has also taught me so much simply by serving as a great model of award-winning, evaluation-busting teaching, of thoughtful research mentoring, and of balanced prioritization of research, teaching, and family. I will certainly try my best to carry forward the "Polk model "as I move into the next stage of my academic career.

I was forewarned by numerous and respectable sources of the possible perils of pursuing an interdisciplinary degree at a large university. These sources clearly knew little of the two departments that ultimately served as the two halves of my degree program.

The EECS Department was an incredibly welcoming place to hang my hat for the first years of my studies and I am grateful to the people and the place for continuing to support me throughout the pursuit of my doctoral degree. I owe special thanks to the AI triumvirate of John Laird, Satinder Baveja and Michael Wellman who lifted my spirits when I managed to *Fail the Qual* and who helped me get through the exam the second time. Thanks also to John Laird for his guidance at the important junctures of my degree, and to Satinder Baveja and Michael Wellman for their teaching mentorship. Lastly, I want to thank Dawn Freysinger for helping me check-the-boxes and weave my way through all the complexities of the interdisciplinary degree process.

The Psychology Department was also an incredibly welcoming place that graciously agreed to provide me a second home when I realized that cognitive psychology would be an important addition to the degree program that I wanted to pursue. In my first meeting, Dave Meyer inquired *Why the #*%$ did you give up a real salary to do a PhD!?* This was the informality and collegiality so typical of the Cognition and Perception Area in which I resided for the second half of my degree. Special thanks to Dave Meyer, for teaching me my first lessons in psychology, for his gruff equanimity ☺, his periodic and always helpful nuggets of wisdom, and for his detailed introductions to the foundations of Rock n' Roll, the "other side" of 50s music, and a part of Country that I never knew. I am very grateful to the department for offering me both academic and financial support. I owe a particular thanks to Lesley Newton, Kathy Hatfield and Scott Paris for enabling me to teach several courses outside my area and experience – opportunities that turned out to be foundational for the research and teaching career that I now plan to pursue. I also want to thank Susan Douglas, Mary Mohrbach, and Nikomi Peltz for being so helpful and so nice in so many ways.

Lastly, I want to thank Stephanie Preston for providing me with a data set that I used for my preliminary research and in this dissertation, and for helping me obtain an additional data set from Antoine Bechara. I also want to thank Tracey Ederer for her diligent contributions to the empirical study reported in Chapter V; it was a pleasure working with her and her involvement made the very solitary endeavor of dissertation research much more fun and rewarding. I also want to thank Antoine Bechara, Josh Weller, and Kyoung-Uk Lee for graciously providing me with data that proved to be critical enabling factors for my dissertation research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER I: INTRODUCTION

Research Context

Decision making, the process of choosing among a set of options, is a fundamental aspect of everyday mental life. Decisions are often made under conditions of uncertainty, when the payoffs are probabilistic and unknown. Should the commuter continue to sit in freeway traffic, or take the next exit and attempt local roads? Should the seller accept the current offer, or reject it for the possibility of a better offer in the future? The study of decision making has been approached from many perspectives – philosophical, behavioral, biological, mathematical, and computational – yet many challenges remain for understanding this important function of higher cognition. Among them, two central challenges are to understand how decision making processes are instantiated computationally in the brain and to reveal and characterize differences in decision making processes across individuals.

Historically, research on decision making in different fields has often proceeded independently. However, in recent years there has been a convergence. Multidisciplinary work in the nascent field of Decision Neuroscience (also referred to as Neuroeconomics) has sought to forge deeper links between brain, behavior and computation, drawing on a large body of behavioral data and theory from psychology and economics, and on an emerging understanding of computational and neural mechanisms as revealed by empirical and theoretical work on the neurobiology of reward learning and motivation. Although this dissertation focuses exclusively on the behavioral and computational approaches to the study of decision making, the three approaches are highly integrated and therefore this dissertation is informed in part by neural data and theory (Figure 1.1). For example, characterization of decision making impaired by neurological damage has contributed to the understanding of decision behavior in healthy individuals, and an intriguingly close correspondence between variables simulated in computational models of reward-learning and their neural counterparts have helped to validate the use of these computational models in testing

mechanistic theories of decision behavior. Furthermore, because of the links between approaches, the research conducted for this dissertation has the potential to contribute to future research using neural approaches. For example, this dissertation focuses on decision making in healthy adults, but its results may have implications for better characterizing and assessing decision making in brain-damaged patients, and the modeling and clustering results may lead to neural predictions that can be tested using imaging methods.

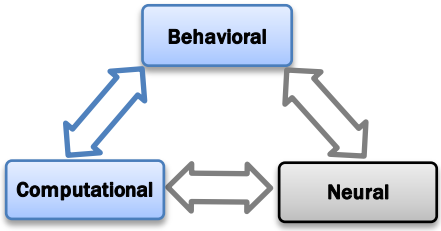

Figure 1.1 Integration of three approaches to the study of decision making. This dissertation focuses on behavioral and computational approaches, but these approaches are informed by neural data and theory.

Decision making is studied normatively as well as descriptively. This dissertation focuses on descriptive accounts of decision making under uncertainty in the context of an important experimental paradigm known as the Iowa Gambling Task (IGT). The overarching aim of this work is to contribute to a better understanding of the important attributes that guide decision making under uncertainty and how individuals differ in these attributes. In this dissertation, I will use the term "individual differences" in two ways that merit definition at the outset. First, I will use the term in referring to a level of analysis in which the focus is on understanding differences in performance within a population using analysis methods performed primarily at the level of individual decision data. I will also use the term somewhat more loosely to refer to differences among *groups* of individuals within a population typically analyzed only in the aggregate.

Background

The Iowa Gambling Task

The Iowa Gambling task is an extensively studied behavioral paradigm that involves decision making under uncertainty. The IGT was designed to simulate the often encountered task of choosing among a set of competing options when payoffs are *a priori*

uncertain and must be learned through experience (Bechara, Damasio, Damasio, & Anderson, 1994). In the IGT, subjects are instructed to choose cards from four decks to maximize the payoffs obtained from their choices across a sequence of trials (Figure 1.2). In the standard version of the task, each chosen card reveals a positive monetary gain, and these gains vary in magnitude. Periodically, cards also include a loss amount which when combined with the gain amount delivers a negative net payoff. The task is typically run for 100 card selections, with the duration not known in advance by participants.

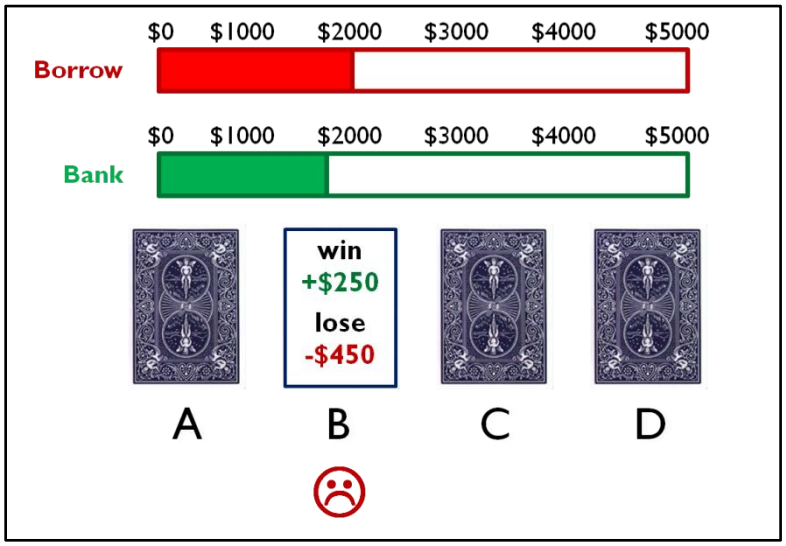

Figure 1.2 The IGT paradigm as administered through a computer. Participants choose cards from four decks. Participants begin the game with an initial loan of $2000. The Bank bar tracks current profits, and the Borrow bar tracks how much money has been borrowed. Wins and losses are indicated visually and auditorily after each choice.

The schedule of payoffs for each deck is fixed by the experimenter. A typical payoff schedule (the *A'B'C'D'* version of the task) is shown in Table 1.1. Decks A' and B' are high gain decks that deliver gains of $80 to $170 on every card, while decks C' and D' are lower gain decks that deliver $40 to $95 on every card. In addition to gain amounts, the decks also deliver periodic losses. On average, decks A' and C' (the *high frequency* decks) deliver a loss on every other card, while decks B' and D' (the *low frequency* decks) on average deliver a loss in one out of ten cards. The magnitudes of these losses vary, with losses in decks A' and B' tending to be larger than in decks C' and D'. The schedule of payoffs is designed such that the larger gains in decks A' and B' are outweighed by the

periodic losses and these decks therefore have negative expected values (each -$72). In contrast, the smaller gains in decks C′ and D′ outweigh the periodic losses and so these decks have positive expected values (each +$32). Decks A′ and B′ are therefore considered the *disadvantageous* or *bad* decks, and C′ and D′ the *advantageous* or *good* decks. Critically, this payoff schedule imposes a tradeoff on decision making: To perform well on the task, decision makers must learn to avoid the choices offering consistently larger gains (A′ and B)′ and instead pursue choices offering consistently smaller gains (C′ and D′) but greater expected value.

Table 1.1 Payoff schedule for the A′B′C′D′ version of the Iowa Gambling Task.

Deck	Deck type	Every-trial gain amounts	Periodic loss amounts	Loss frequency	Expected value
A′	Bad	$80 to $170	–$150 to –$350	50%	–$72
B′	Bad	$80 to $170	–$1250 to -$2500	10%	–$72
C′	Good	$40 to $95	–$25 to –$75	50%	+$32
D′	Good	$40 to $95	–$250 to -$375	10%	+$32

Notes. Decks A′ and B′ are the disadvantageous decks (bad decks) that deliver larger gains on every card, but have periodic losses that outweigh the gains. These bad decks deliver a negative expected value (-$72). Decks C′ and D′ are advantageous decks (good decks) that deliver smaller gains on every card, but these gains outweigh the periodic losses yielding a positive expected value (+$32). The decks also differ in the frequency of their losses (Bechara, Tranel, & Damasio, 2000).

Analyzing Performance in the IGT

Empirical research on the IGT has identified a robust set of performance characteristics (see B. D. Dunn, Dalgleish, & Lawrence, 2006 for a review) which have been shown to hold under a wide range of variations in the task, for example: real vs. faux monetary payoffs (Bechara, Tranel, et al., 2000), inversion of the gain/loss schedules (Bechara, Damasio, & Damasio, 2000), manual versus computer administration (Bechara, et al., 1994), and payoff schedules with fixed (van den Bos, Houx, & Spruijt, 2006) versus diverging differences between gains and losses (Maia & McClelland, 2004). The primary dependent variable upon which IGT performance is typically assessed is the total percentage of cards selected from the two good decks *(%Good*), or equivalently the difference between the mean number of good and bad selections (i.e., [C′+D′] - [A′+B′] where A′,B′,C′,D′ are either the total number or the percent of cards selected from each deck). A participant who performs advantageously selects more cards from the good

decks (C′ and D′) than the bad decks (A′ and B′) over 100 trials. In aggregate, a population of healthy subjects performs advantageously, typically selecting approximately 60-70% of their cards from the good decks (Figure 1.3A). Clinical populations with decision making deficits show an opposite pattern of performance, selecting in aggregate more cards from the bad decks (Figure 1.3B).

One logical problem with the use of *%Good* as a dependent measure is that early in the task participants experience few losses due to the way the fixed payoff schedule is designed (Maia & McClelland, 2004). As a result, the expected values of the "bad" decks *as experienced* by participants actually tend to be positive and larger than for the "good" decks. Maia and McClelland (Bechara, et al., 1994; Damasio, 1994) proposed an improved measure that tracks the percentage of cards selected by participants from the two decks that, on any given trial, have the highest *experienced expected value* up to that trial (*%EEV)*. By this measure, advantageous performance is defined in terms of deck values participants actually experience rather than the good/bad labels assigned to the decks a *priori* by the experimenters. Practically, a divergence between *%Good* and *%EEV* is manifested primarily in the early trials, and only rarely in later trials. Aggregate IGT performance measured according to *%EEV* therefore shows a similar overall pattern of performance as measured by *%Good*.

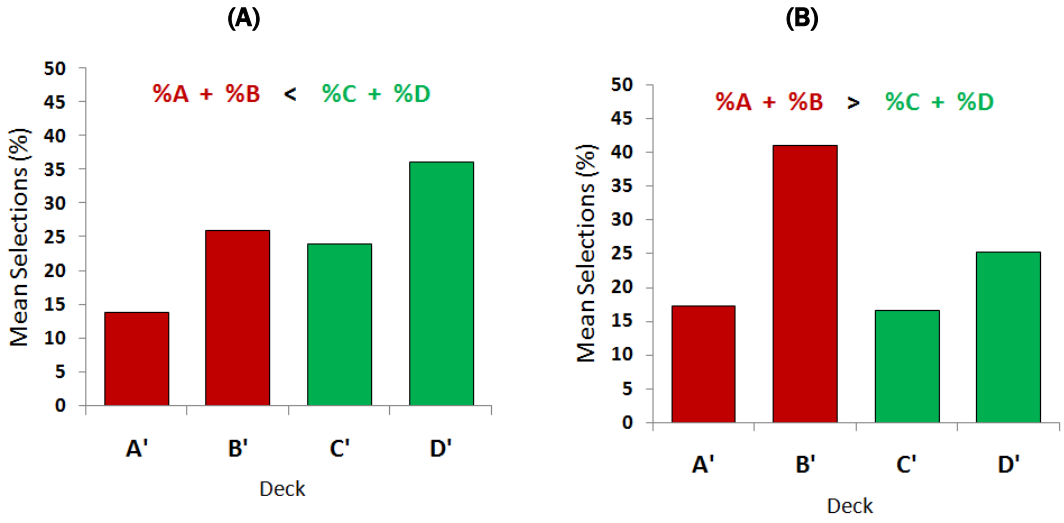

Figure 1.3 Typical patterns of IGT card selections. (A) 844 healthy adults and (B) 11 patients with damage to ventromedial prefrontal cortices. In aggregate, healthy participants select more cards from the good decks (C′ and D′) than the bad decks (A′ and B′), while the patients exhibit the opposite pattern of performance. Percentages for each deck were pooled over trials and participants. Source: (A) multiple IGT data sets obtained for the purposes of this dissertation (see Chapter IV for a detailed description of these data); (B) data courtesy of Dr. Antoine Bechara, University of Southern California.

Performance on the IGT has also been analyzed in terms of the time course of selections, either on a trial-by-trial basis, or more often aggregated in five blocks of 20 trials. This temporal transition is shown in Figure 1.4 for a population of 844 healthy participants. In the first block of trials, healthy participants typically sample from each of the decks and exhibit an early preference for the bad decks that offer consistently larger gains. However, as participants learn more about the payoffs, their choices shift to the good two decks that offer positive expected value.

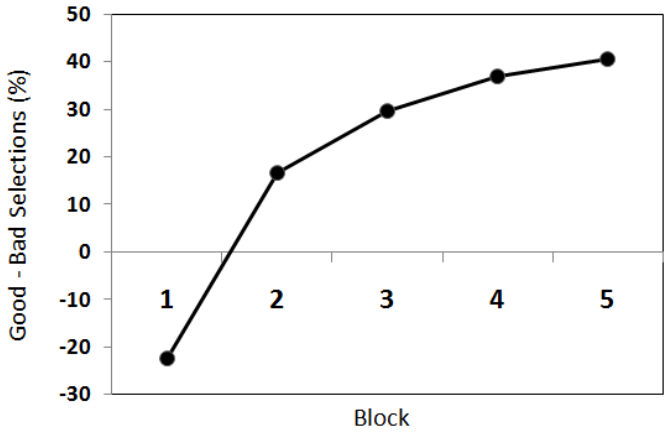

Figure 1.4. Trajectory of IGT choice behavior for 844 healthy participants. The plot shows the mean percentage of cards selected from the two good decks minus the two bad decks across five blocks (each of 20 trials). Source: Multiple IGT data sets obtained for the purposes of this dissertation (see Chapter IV for a description of these data).

The Use of the IGT in Research

The IGT was originally developed as an experimental paradigm diagnostic of the types of real-world decision making deficits exhibited by patients with damage to ventromedial prefrontal cortex who exhibit the pattern of performance shown in Figure 1.2B (Bechara, Damasio, & Damasio, 2003; Bechara, Damasio, Damasio, & Lee, 1999). Its use has subsequently been extended to the study of: (i) neuropsychological conditions beyond ventromedial prefrontal cortex, for example the amygdala (Campbell, Stout, & Finn, 2004; Mimura, Oeda, & Kawamura, 2006; Pagonabarraga, et al., 2007; Stout, Rodawalt, & Siemers, 2001) and basal ganglia (B. D. Dunn, et al., 2006 provides a summary as of 2005); (ii) a wide range of psychopathologies including substance abuse and dependency, pathological gambling, obsessive-compulsive disorder, schizophrenia, attention deficit disorders, eating disorders, impulse and aggression disorders, and psychopathy (Bolla, Eldreth, Matochik, & Cadet, 2004; Desmeules, Bechara, & Dube, 2008; Franken & Muris, 2005; Preston, Buchanan, Stansfield, & Bechara, 2007; Reavis & Overman, 2001; Suhr & Tsanadis, 2006; Sweitzer, Allen, & Kaut, 2008; van Honk, Schutter, Hermans, & Putman, 2003; Zermatten, Van der Linden, d'Acremont, Jermann, & Bechara, 2005); (iii) the link between decision making and personality traits, affective states, and other individual differences among healthy individuals (Crone, Bunge, Latenstein, & van der Molen, 2005; Crone & van der Molen, 2004, 2007; Garon & Moore,

2004, 2007; Hooper, Luciana, Conklin, & Yarger, 2004; Huizenga, Crone, & Jansen, 2007; Kerr & Zelazo, 2004); (iv) the development of decision making and executive function across the lifespan (Bolla, et al., 2004; Bolla, Eldreth, Matochik, & Cadet, 2005; Fukui, Murai, Fukuyama, Hayashi, & Hanakawa, 2005; Lawrence, Jollant, O'Daly, Zelaya, & Phillips, 2009; Oya, et al., 2005; Schutter, de Haan, & van Honk, 2004), and (v) the neural correlates of decision making under risk and uncertainty (Bechara, 2007).

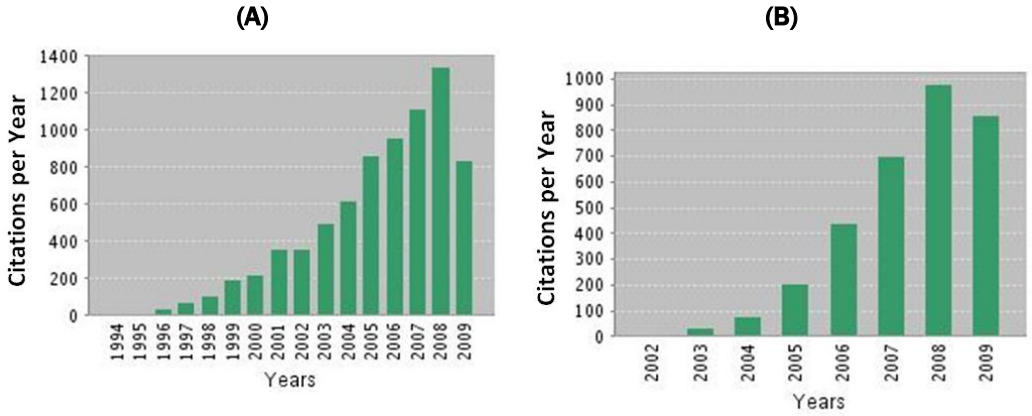

Figure 1.5. The IGT continues to be a widely used and studied experimental paradigm. (A) Number of yearly citations of the core set of IGT papers. (B) Number of yearly citations matching the keyword "Iowa Gambling Task". Source: Journal Citation Reports®, ISI Web of Knowledge: Thomson Reuters, Inc.

The IGT continues to be an important task used for basic research on decision making in both healthy and clinical populations, as evidenced by the fact that it has been administered and/or cited in thousands of published studies (Figure 1.5). The IGT is also a widely used tool for clinical assessment and its recent commercial availability is likely to further contribute to its continued use for this purpose (for example, Xiaohua & Illhoi, 2004).

Aims and Methods

Given the important role of the IGT in the study of decision making, a comprehensive conceptualization of behavior in this task is necessary as a foundation for inference. This is particularly true for studies of clinical populations where results have implications for how we conceptualize neurological disorders and psychological disturbances in decision making. A comprehensive conceptualization should include an accurate framework for understanding important attributes of performance and should

identify and characterize how performance differs across individuals, if fundamental differences exist.

There are reasons to believe that the current conceptualization of the IGT may be incomplete. These reasons will be discussed in the chapters that follow, but in brief these include: (i) the presence of large minorities of healthy participants who perform disadvantageously, yet show no real-life decision making impairments; (ii) unreliable findings in studies that have tested the association between IGT performance and demographic, trait, and cognitive measures, (iii) high variances in group-averaged model parameters fit to decision data; and lastly (iv) a set of important assumptions underlying IGT analysis methods that have not yet been tested. If the current conceptualization of the task is incomplete, this would have important implications for the measures, methods of inference, and approaches to clinical assessment that are currently being used.

The primary research objective of this dissertation was to investigate performance in the IGT at a lower-level of analysis and using more thorough methods than have previously been used in analyzing this task. The dissertation consists of a set of studies that reflect the interdisciplinary nature of the doctoral degree that I am defending. The dissertation consists primarily of three primary objectives and three methods (Figure 1.6). First, I used computational models to try to better understand the IGT in terms of decision attributes and mechanisms (Chapter II). These models were independently motivated by theory and data on reinforcement learning systems in brain that have been previously been shown to provide a good account of related decision tasks in humans and animals. The aims of this first study were to (i) provide a better account of the IGT than provided by the currently accepted model of the task and (ii) to test the efficacy of the reinforcement learning framework as an account of behavior in the IGT. The second major objective was to test for the presence of important individual differences in performance in the IGT, or alternatively confirm that the task is well-captured through population-level analysis (Chapter III and IV). To pursue this objective, I used unsupervised clustering methods drawn from the fields of machine learning and data mining and widely used in the analysis of gene expression data (Bechara, et al., 1994; Bechara, Damasio, Tranel, & Damasio, 1997; Bechara, Tranel, Damasio, & Damasio, 1996). Having found robust individual differences in the task, I then pursued a third and final

objective which was to try to further characterize differences in the IGT in terms of more fundamental measures independent of performance in the IGT (Chapter V). To pursue this aim, I conducted an empirical study in which I collected IGT data from a set of healthy participants concurrent with collection of demographic, trait, and cognitive measures.

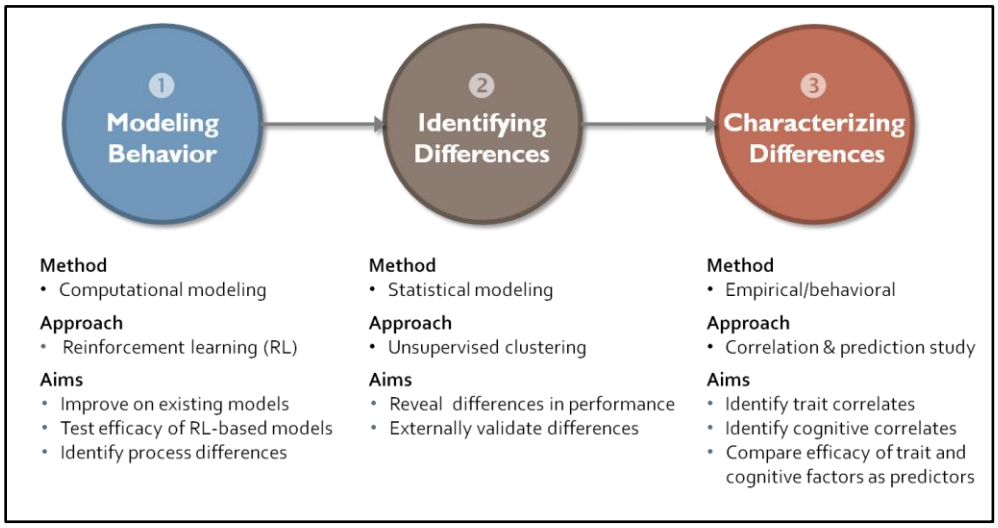

Figure 1.6 Overview of the primary objectives and methods undertaken in the dissertation.

CHAPTER II: MODELING DECISION MAKING

Introduction

Motivation

There is a large body of empirical data on the IGT, but at present there is no explicit, comprehensive theory of performance in this task. Given the ongoing use of the IGT in the study of decision making, additional theoretical work is merited. The two aims of this study were to investigate the ability of a set of computational reward-learning models to (i) capture the core phenomena associated with the task, and (ii) to account for variability in the performance of individual participants.

It has been demonstrated that damage to ventromedial and orbitofrontal regions of prefrontal cortex (vmPFC) can lead to marked impairments in real-world decision making and to impaired performance in the IGT (for example, Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Rolls, 1996, 2004; Schultz, 1998, 2006; Schultz, Tremblay, & Hollerman, 2000; Thut, et al., 1997; Tremblay & Schultz, 1999). Interestingly, these same brain regions are among those also known to be involved in reward-based learning in both animals and humans (for example, Kakade & Dayan, 2002; Montague, Dayan, Person, & Sejnowski, 1995; Niv, Daw, & Dayan, 2006; J. P. O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Schultz, Dayan, & Montague, 1997) This commonality in the brain areas involved in reinforcement learning processes and in impaired IGT performance suggests that theories of reward-based learning developed in behavioral neuroscience might be good candidates as theories of performance in the IGT.

In the past decade, theoretical work on reward-based learning in the brain has made contact with computational work developed relatively independently by learning theorists in computer science. This productive connection has led to explicit computational models of reward learning in the brain. Critically, these reinforcement learning (RL) models have been validated against neural data in both animals and humans (Bechara, Damasio, et al., 2000; Damasio, 1996). Taken together, the established link between IGT performance and neural mechanisms of reward learning, and the

availability of neurobiologically-grounded models of these mechanisms motivates the application of computational RL models to the study of performance in the IGT, as was done in the research reported here.

While motivated independently, there is an important connection between the RL-based modeling work undertaken here, and the principal theory of the IGT put forth by the original developers of the task in what is known as the Somatic Marker Hypothesis or SMH (See B. D. Dunn, et al., 2006 for a detailed review of the SMH). At the core of the SMH is the assertion that in choosing among options of uncertain value, decision processes are guided by affective biasing signals that originate either in the body or cortical representations of the body (Colombetti, 2008; B. D. Dunn, et al., 2006). The act of experiencing rewarding or punishing stimuli generates affective responses that are stored as somatic states or "as-if" representations of these states in somatosensory cortex. These stored somatic states, or *somatic markers*, are useful future indicators of stimulus value that can serve to guide decision making. Although based on a substantial body of behavioral and physiological data, the SMH is a verbal theory, and one which has not been made sufficiently explicit such that it can either be falsified, or distinguished from more explicit accounts that might be developed (Busemeyer & Stout, 2002). However, two of the core claims of the SMH are fully consistent with the core premises of reinforcement learning theory and it is therefore not unreasonable to consider RL models of the IGT as partial instantiations of the SMH. The core claims shared by these theories, stated in terms of the SMH are: (i) that immediate reward-based experience generates longer-term markers of value, and (ii) that these markers serve as signals that guide decision making. The SMH, however, makes the further claim that these markers originate from somatic representations of reward, a claim not shared by RL theory which makes no assumptions about the internal representation of reward.

Prior Work

Previous computational studies of the IGT have used primarily both heuristic and formal models developed in the fields of decision theory and mathematical psychology. In this initial work, three alternative models (Heuristic Strategy Switching, Bayesian Expected Utility, and Expectancy-Valence) were fit to IGT data and compared using model selection methods (Sutton & Barto, 1998). The expectancy-valence model was shown to best fit the IGT data. Interestingly, the expectancy-valence model is

isomorphic to the single-state, Q-learning model that has been well developed in the computational reinforcement learning literature (Busemeyer & Stout, 2002; Kalidindi & Bowman, 2007; Lane, Yechiam, & Busemeyer, 2006; Stout, Busemeyer, Bechara, & Lin, 2002; Stout, Busemeyer, Lin, Grant, & Bonson, 2004; Yechiam, Busemeyer, Stout, & Bechara, 2005), but this existing modeling work on the IGT has made little contact with this literature. Since its initial publication, the expectancy-valence model has become the accepted model of the IGT and has been put to use in clinical settings to help characterize impaired decision making due to neurological damage and disorder as well as to a wide range of psychopathologies. In clinical use, the expectancy-valence model has been fit to clinical populations and the population-averaged parameters of the model were then compared to the parameters obtained from healthy controls. Differences in the parameters across populations were used as the basis for inference in characterizing differences in decision making (for example, Bechara, et al., 1997).

The expectancy-valence model represents an important first step in offering an explicit computational framework for studying the IGT. Furthermore, work using this model has demonstrated how an explicit computational model of the task might be used as a way to study and characterize impaired decision making. However, there are important limitations in this prior work that motivate, in part, the aims of the present study. First, the expectancy-valence model was selected from among a set of three models. One of these models was a Bayesian decision model. Given the well-documented departure of human decision making from the normative accounts offered by Bayesian theory, one would not necessarily expect this candidate model to do well in explaining the IGT. The second model, the Heuristic Strategy-Switching model was not independently motivated, but rather it instantiated directly in its equations the core behavioral phenomena to be modeled, namely an initial preference for the bad decks, followed by a switch to the good decks. While few (if any) model comparisons studies are able to explore the full space of possible models relevant to modeling a psychological task, that the expectancy-valence model was selected from among a set of three models does not provide strong support for its validity, and certainly suggests a broader search for a better model might be productive. As a standard reinforcement learning model, the expectancy-valence model is among a class of models with many variants that have been

explored by learning theorists. Little or no work has been done to determine whether other members of this class provide a better account of IGT performance.

A second important limitation is that the assumptions underlying the use of the expectancy-valence model as a tool for inference across populations have not been well tested. While models have been fit to individual decision data, model selection and inference have been done using population-averaged parameters. Implicit in population-level selection and inference is an assumption that a population is a reasonable representation for some, if not most individuals. This is an important assumption that does not necessarily hold. As an extreme example, an average computed from a population comprised of two subsets performing at opposite extremes on a dependent measure, contains no individual that is similar in performance to the population average. The validity of this assumption has great import given that the expectancy-valence model is being used for clinical assessment. Yet, to the author's knowledge no prior work has challenged this assumption and there are several reasons to believe it does not hold. The IGT is a complex task as evidenced by the robust finding that a large number of normal participants (typically 20-40%) routinely fail to perform the task successfully (Bechara, et al., 1994; Bechara, et al., 1997; Bechara, et al., 2001; Bechara, et al., 1996; Maia & McClelland, 2004; Tomb, Hauser, Deldin, & Caramazza, 2002; Turnbull, Berry, & Bowman, 2003). While this finding might be explained by extra-task factors (e.g. motivation, attention) it might instead result from the presence of more fundamental differences in the way some individuals perform the task. The high variance in population-averaged model parameters reported using the expectancy-valence model also suggest the possibility that this model (and possibly the entire class of models) may not accurately characterize decision making behavior for a large subset of participants.

In combination, these limitations motivate the aims of the present study: (i) to use a base RL model to replicate the type of group-averaged results reported in prior modeling work, (ii) to compare the base model to a diverse set of model variants to identify possible individual differences in decision making, and (iii) to analyze the results across models and individuals to assess the overall efficacy of this class of RL models in capturing performance in the IGT.

Empirical Phenomena to be Modeled

As discussed in the introductory chapter, the primary dependent variable upon which IGT performance is assessed is the total percentage of cards selected from the good decks (c.f., Figures 1.3A and 1.4). In addition to this aggregate measure, a robust set of behavioral phenomena have been reported in the IGT literature (Sutton & Barto, 1998) and can be identified through inspection of individual performance data. The selection history of typical participant exhibiting advantageous performance and the important behavioral phenomena associated with this performance are summarized in Table 2.1 and shown in Figure 2.1.

Table 2.1 Behavioral phenomena in the IGT.

Phenomena	Description
(1) Exploration	In the early trials participants typically sample from all four decks, often selecting one card from each deck in turn, or selecting several cards from each deck before sampling from another.
(2) Early Bias for Bad Decks	In the first twenty trials, participants typically choose more cards from the bad decks.
(3) Shift to Good Decks	Typically after twenty trials, participants that ultimately succeed in the task begin choosing more cards from the good decks. Advantageous selection continues until the end of the experiment.
(4) Resampling	In the last 10-20 trials, many participants make occasional selections from the bad decks.

Notes. The number listed with each phenomenon are associated with the numbers labeling the plots in Figures 2.1 and 2.2.

Participants begin the experiment with the knowledge that cards will deliver gains and losses, but no knowledge about the payoff schedule (c.f., Appendix D for example of IGT instructions). In the initial trials they typically sample from each of the decks to learn more about the nature of these payoffs (phenomenon 1). In subsequent trials, participants exhibit an early preference for the bad decks, typically deck B (phenomenon 2). As the experiment progresses participants' selections shift to the good decks (phenomenon 3) and in the later trials they make occasional selections from the bad decks (phenomenon 4). The advantageous participant in Figure 2.1 shows each of these phenomena and it is evident from the figure that this participant made many selections from the deck that had the highest experienced expected values on a given trial (pink markers).

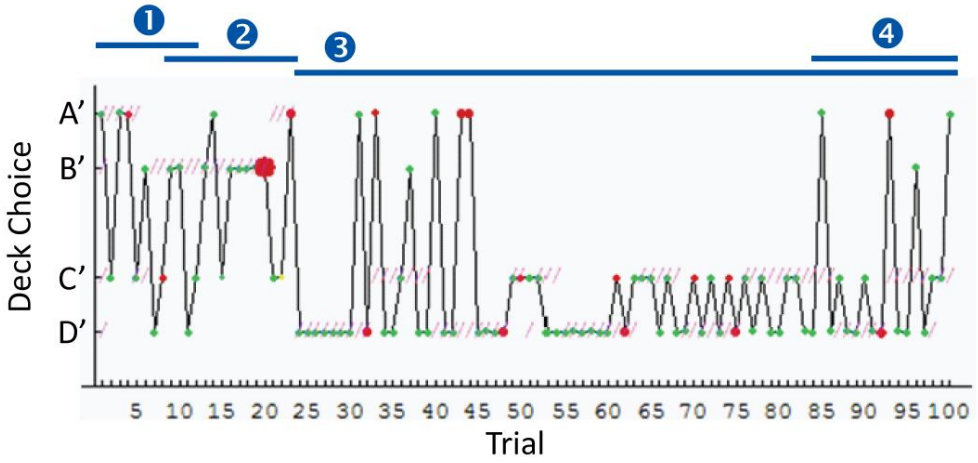

Figure 2.1 Patterns of deck selections for a healthy participant who performed advantageously. The core phenomena are numbered and described in Table 2.1. Red dots indicate loss trials, green dots gain-only trials. The size of the dot indicates the size of the gain or loss. The pink markers indicate the deck with highest experienced expected value on a each trial. Source: Data set described in methods section of this chapter.

In contrast, disadvantageous participants within a healthy population show a different pattern of selections (Figure 2.2). Although these participants exhibit the same initial exploration and preference for the bad decks, they do not show a shift from the bad decks to the good decks and instead exhibit a continuing preference for the bad decks. The disadvantageous participant shown in the figure made few choices from the decks with the highest (pink markers) and second highest (orange markers) experienced expected values on a given trial, suggesting the possibility that other attributes of choice were guiding behavior.

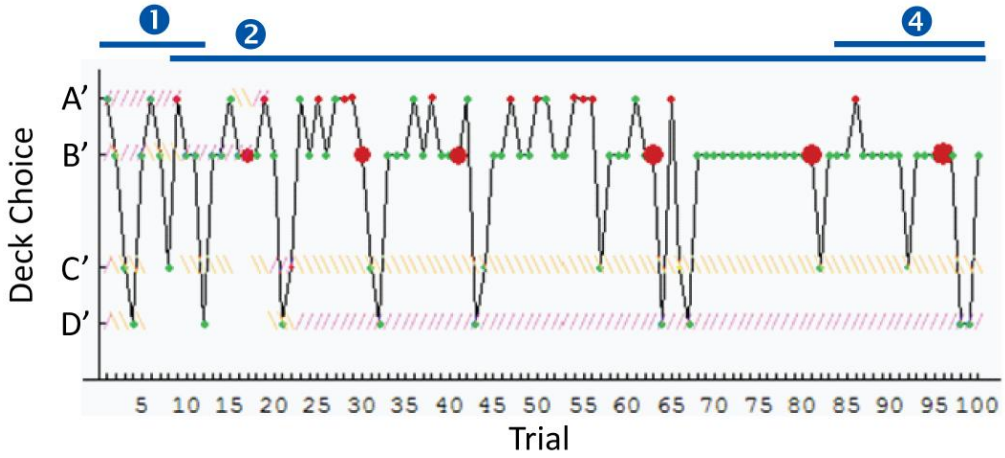

Figure 2.2 Patterns of deck selections for a healthy participant who performed disadvantageously. The core phenomena are numbered and described in Table 2.1. Red dots indicate loss trials, green dots gain-only trials. The size of the dot indicates the size of the gain or loss. The pink markers indicate the deck with highest experienced expected value, and the orange markers indicate the deck with the second highest experienced expected value on each trial. This participant does not show the a transition from the bad decks to the good decks (phonomenon3). Source: Data set described in methods section of this chapter.

While there is variability in trial-by-trial patterns of deck selections across participants, in aggregate healthy participants exhibit a robust shift from an initial preference for the bad decks to a preference for the good decks (c.f., Figures 1.3A and 1.4).

Models

Formally, the IGT can be modeled as a reinforcement learning (RL) problem in which participants must learn, via experienced rewards, the relative values of the four decks. It is isomorphic to the n-arm bandit problem that has been extensively analyzed in the machine learning literature (Dayan & Abbott, 2001; Montague, et al., 1995). It is also isomorphic to the classic bee foraging problem addressed in the behavioral and neuroscience literature (Dayan & Abbott, 2001). In the IGT, because there is no delay between the choice of deck and the receipt of reward the task falls within the static-action class of reinforcement learning problems (Sutton & Barto, 1998). The IGT may also be considered as a non-associative reinforcement learning problem (Luce, 1959, 1977) under the assumption that participants learn to select advantageously from a single state, and therefore learned actions need not be associated with multiple states. The IGT, could, however, be modeled as an associative problem where the representation of state

might include a metric of current profitability and/or attributes that capture aspects of a participant's declarative knowledge. As a starting point, I pursued the non-associative approach and I leave the associative possibility for future work. Action value models (also known as delta-rule models) are a standard approach to modeling non-associative, static-action RL problems and this was the approach pursued in this study.

The Base Model

The base model used in this study had three components: (i) a *learning component* responsible for storing and updating estimates of the value of each deck based upon experienced payoffs, (ii) a *selection component* that chooses cards from the decks based on their estimated values, and (iii) a *reward function* that translates experienced monetary payoffs into an internal representation of reward.

Learning component

The learning component of the base model served the role of applying current experience to update previously stored knowledge. In learning to choose cards advantageously, the model assumed that participants have a representation of value (V_A, V_B, V_C, V_D) for each deck (A, B, C, D) and that these values are updated based on the payoffs experienced from each deck during the task. The values were initialized to zero and updated on each trial according to the following learning equations, known as the delta rule:

$$V_i(t+1) = (1-\alpha) \cdot V_i(t) + \alpha \cdot r(t) : i \in [A, B, C, D] \quad (1a)$$

$$V_i(t+1) = V_i(t) + \alpha \cdot \big(r(t) - V_i(t)\big) : i \in [A, B, C, D] \quad (1b)$$

In the learning component, *r(t)* was the internal reward experienced by the participant after receiving the monetary payoff from the deck selected on trial t. The form of equation (1a) makes clear that α, the learning rate, governs the relative influence of the current value estimate $V_i(t)$ and the current reward *r(t)* on the updated value $V_i(t+1)$. Alternatively, the form of equation (1b) highlights the fact that values are updated based on an error term $r(t)$-$V_i(t)$ that represents the difference between the actual reward *r(t)* experienced from deck *i* on trial *t*, and the currently stored estimate of that value as represented by $V_i(t)$.

Selection component

The selection component of the model produced deck choices based on the stored values for each deck. The base model assumed that on each trial, participants select

probabilistically from among the decks based on the current estimated values stored with each deck. The assumption of probabilistic choice was mathematically codified in Luce's well-known Choice Axiom which relates choice probabilistically to relative value (Herrnstein, 1961). Probabilistic choice was also demonstrated in instrumental conditioning tasks by Herrnstein who observed that the frequency of responses among a set of options was closely matched to the frequency of rewards delivered by those responses (Lau & Glimcher, 2005). It has also been shown that response frequencies also match the magnitudes of the rewards associated with each response (J. P. O'Doherty, et al., 2004). The applicability of the probabilistic choice rules to human behavior has also been demonstrated more recently in neuroimaging studies of human participants. (Corrado, Sugrue, Seung, & Newsome, 2005; Daw & Doya, 2006; Lau & Glimcher, 2005).

Additional work has clarified the functional form of matching-based selection, and found that a softmax selection rule (discussed below) best fits behavioral choice data (Sutton & Barto, 1998). Interestingly, the softmax rule is also a predominant model of action selection in the machine learning literature (Daw & Doya, 2006). The softmax rule is value-sensitive method for implementing probabilistic selection because the relative choice probabilities are proportional to the relative value estimates associated with each deck. This rule was recently tested in a human choice task against an undirected probabilistic rule (ε-greedy) and against a more sophisticated softmax rule that gives greater weight to actions for which selection might resolve outstanding uncertainty (Sutton & Barto, 1998). The base softmax rule was found to provide the best fit to the data.

In the present study, the probability $P_d(t+1)$ that a participant selects deck d on trial $t+1$ was computed according to the following softmax rule (for example, Tremblay & Schultz, 1999)

$$P_d(t+1) = \frac{e^{\theta \cdot V_d(t)}}{\sum_j e^{\theta \cdot V_j(t)}} \; : \; j, d \in [A, B, C, D] \qquad (2)$$

In this selection equation, the sensitivity parameter θ determines the degree to which differences in the deck values V_i are transformed into differences in selection probabilities. When θ is zero, selection probabilities are uniform regardless of the value estimates and therefore selection is random. Large values of θ lead to large differences in

selection probabilities, and as θ approaches infinity selection becomes deterministic, with the highest valued deck always selected.

Reward function

The reward function in the model transformed monetary payoffs (stimuli) experienced by the participant during the task into an internal scalar representation of reward *r(t)*. In the base model, trial rewards were computed according to the following equation:

$$r(t) = G \cdot gain(t) + L \cdot loss(t) \quad (3)$$

In the base model, the rewards represented the net monetary payoff obtained on a given trial, with the parameters *G* and *L* allowing for both relative weighting of the contribution of gains and losses to the reward as well as absolute weighting of the levels of reward experienced by a participant.

Model Rationale

While the specific form of the base model did not depart from standard single-state Q-learning model in the RL literature, it is nevertheless important to note that the choice of this model was grounded in both empirical and theoretical considerations. These considerations are outlined below.

Attention to gains and losses

There is evidence that the reward value of stimuli are represented in vmPFC, the same brain area associated with patient deficits in the IGT (J. O'Doherty, Rolls, Francis, Bowtell, & McGlone, 2001). In addition, there is evidence that representations of rewarding and punishing stimuli in the human brain are dissociable. In a recent human neuroimaging study, medial areas of OFC were found to be activated by monetary rewards while more lateral areas were responsive to monetary losses (Kahneman & Tversky, 1979) The magnitude of neural responses in these two areas were also found to be correlated with the magnitude of the monetary payoffs. That gains and losses are weighed differently is also consistent with the well-known finding in behavioral economics that "losses loom larger than gains" (Daw, et al., 2006). The base model provided for this possibility through the use of independent gain and loss parameters as a way to investigate the possibility that participants performing the IGT place differing absolute and relative value on the monetary gains and losses obtained in the task.

Influence of past rewards on learned values

In the original set of IGT studies, vmPFC patients exhibited a preference for the decks that deliver higher immediate gains, with a putative myopia for the longer-term net consequences that result from the pursuit of this preference. A subset of healthy control participants also exhibit disadvantageous performance. One algorithmic explanation for impaired performance is the possibility that participants (and patients) differ in the degree to which current payoffs are able to influence value updates. The base model provided for this possibility via the learning rate parameter α. Larger values of α increase the influence of current payoffs on the updating of values, while smaller values of α have an opposite effect.

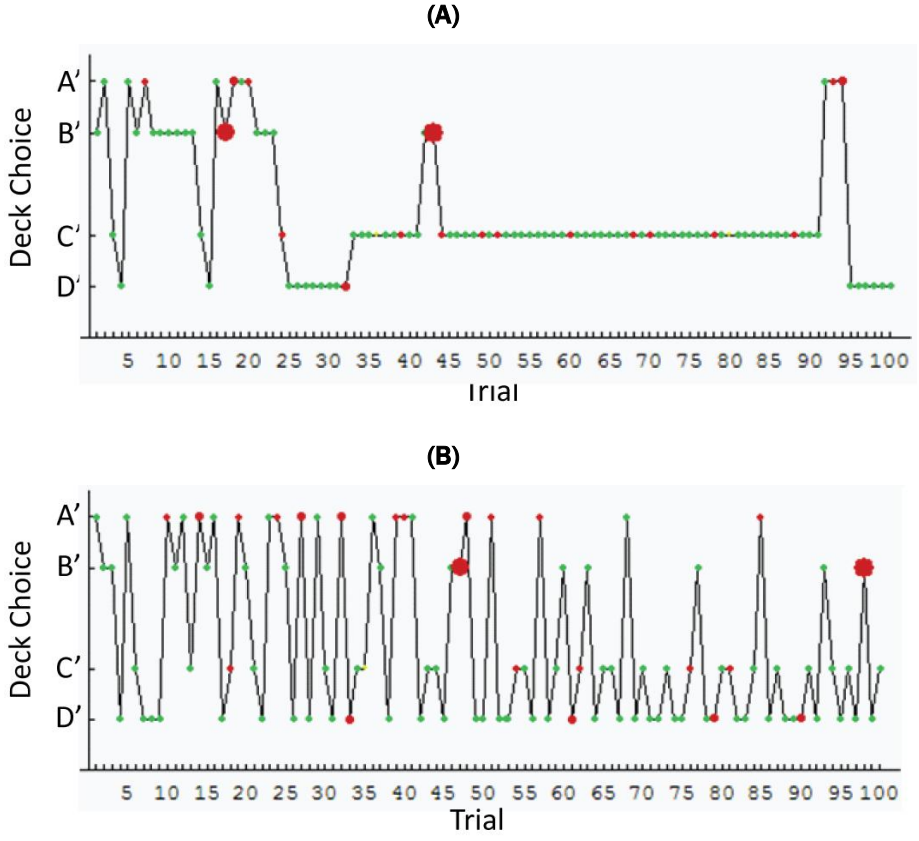

Figure 2.3 The IGT selection histories for two healthy participants. Both participants performed advantageously, but the selections of the participant shown in (A) appear more deterministic than those of the participant shown in (B). This raises the possibility that participants may differ in the sensitivity of choice based on learned value estimates. Source: Data set described in methods section of this chapter.

Control of probabilistic deck selections

Reinforcement learning problems typically involve a tradeoff between exploring options for the benefit of improving learned value estimates and exploiting these estimates to improve performance. A recent neuroimaging study localized these two competing functions to different neural substrates (Sutton & Barto, 1998). It is therefore plausible that participants may differ significantly in the degree to which they explore versus exploit. The base model allowed for these differences via the sensitivity parameter θ in the selection equation. Figure 2.3 shows an example of two participants who both performed the IGT advantageously, but who seemingly differed in the degree of determinism in their choices.

Model Variants

For the purposes of identifying the algorithmic assumptions that might better fit IGT participant data, I explored a range of variations of the base model, each of which was motivated theoretically and empirically. The models varied along five dimensions: (i) the functional form of the learning rate (exponential vs. simple averaging), (ii) the nature of action selection (softmax vs. pursuit), (iii) the nature of prediction error (delta-rule versus reinforcement comparison), (iv) the presence or absence of value decay, and (v) the functional form of the reward function. This section presents each of these model variants and their motivations. The variations in the base model were considered one-at-a-time to determine the extent to which each might capture unique aspects of individual decision behavior. I believe this approach was appropriate given the aims of the study which were to investigate the efficacy of RL models in capturing behavior in the IGT and to identify possible differences in decision behavior across individuals. A full factorial investigation of model variants would be appropriate in seeking a "best" model from among this class, and this is left for future work.

Simple versus exponential averaging

In Equation (1), the learning rate parameter α determined the relative influence of current and past rewards on the updating of value estimates. In the base model, α was a constant. A constant α leads to an exponential weighting of past rewards, with more recent rewards having exponentially greater influence on learned values than more remote rewards. In a variant model, α varied inversely with the number of times a deck had been selected. This produced learned values for each deck that were simple

arithmetic averages of all rewards experienced. For non-stationary and effectively non-stationary problems, the exponential form has been shown to be advantageous (J. P. O'Doherty, Critchley, Deichmann, & Dolan, 2003), but given lack of knowledge about how participants perceive the sequence of payoffs in the IGT, both the simple- and exponential-average methods were utilized.

Independently stored action probabilities

In the base model, value estimates were stored and updated on every trial. Selection probabilities were computed on each trial using these stored values. There is some evidence that selection processes and value learning and storage processes might be supported by different neural substrates (Sutton & Barto, 1998). Therefore, selection probabilities themselves may be independently stored and subject to a different course of learning than the action values upon which they are derived. This possibility was captured by a variant model (Pursuit model) that utilized a pursuit method of selection (for example, Kahneman & Tversky, 2000; Tversky & Kahneman, 1981). In this model, a selection probability P_j was stored for each deck, and these probabilities were updated on each trial according to the following equations.

$$P_{j^*}(t+1) = P_{j^*}(t) + \beta \cdot \left(1 - P_{j^*}(t)\right) \; : j^* = argmax_j\left(V_j\right),\; j \in [A, B, C, D] \qquad (4a)$$

$$P_j(t+1) = P_j(t) + \beta \cdot \left(0 - P_j(t)\right) \quad \forall j \neq j^*,\; j \in [A, B, C, D] \qquad (4b)$$

In the Pursuit model, value estimates were computed and stored as in the base model, but on every trial the probability of the choice with the highest value was moved towards one, and the probabilities all other choices were moved towards zero. The parameter β determined the rate at which stored probabilities were updated. The effect of this change to the model was that frequent selection of a deck made this deck more likely to be chosen again in the future, a separate influence on choice acting in concert with the influence of the stored value estimates.

Learning using a reference reward

In the Base model, the updates to stored values were made based on the prediction error computed as the difference between the current reward r(t) and the current value estimate $V_i(t)$ (see Equation 1b). Reference effects, the dependence of choice on a context-specific reference point, are well documented in the psychological literature on decision making (Sutton & Barto, 1998). Therefore, I investigated the possibility that

deck selections in the IGT might be affected by a reference level of reward, i.e. premised upon the establishment of a reference level of payoff upon which each selection is compared. A natural way to implement this possibility computationally was via a Reinforcement Comparison model (see Shizgal, 1997 for a review). In this model, I modified Equation (1b) so that prediction error was based on the difference between the current reward *r(t)* and a reference reward $\bar{r}(t)$ which was learned across trials:

$$V_i(t+1) = V_i(t) + \alpha \cdot \big(r(t) - \bar{r}(t)\big) : i \in [A, B, C, D] \quad (5a)$$

$$\bar{r}(t+1) = \bar{r}(t) + \beta \cdot \big(r(t) - \bar{r}(t)\big) \quad (5b)$$

The reference reward was updated on every trial, based on the rewards experienced from all decks. The parameter β governed the rate at which the reference reward was updated.

Limitations in the maintenance of learned values

In the Base model, the only value updated on each trial was the value of the selected deck. Implicit in this model was the assumption that previously learned values for all other decks are faithfully maintained. The task demands of the IGT are substantial enough that 20% of healthy participants typically fail to perform advantageously. It is therefore possible that as a result of task demands and/or noise in learning processes, not all participants are able to maintain accurate value estimates for decks which have gone unselected over multiple trials. I instantiated this possibility computationally in a variant model (Decay Model) in which on every trial that a deck went unselected, the value estimate for that deck decayed towards zero.

$$V_{i^*}(t+1) = V_{i^*}(t) + \alpha \cdot \big(r(t) - V_{i^*}(t)\big) : \; i \in [A, B, C, D] \quad (6a)$$

$$V_i(t+1) = \beta \cdot V_i(t) \; \forall i \neq i^* \quad (6b)$$

In this model, value updates for the selected deck i^* on each trial (Equation 6a) were identical to the base model (Equation 1a). The value of unselected decks decayed by the fraction β which as fixed at 0.8 (i.e. a constant decay rate of 20%).

Individual differences in the nature of reward

Although the reward function is a critical component in the computational RL framework, it has received very little research attention. The mapping between stimuli and environmental conditions and internal representation of rewards is considered domain-specific, and as such have typically thought to lie outside the RL framework

itself. The standard reward function typically produces scalar values based on magnitude quantities experienced in the environment (e.g., amount of money, hedonic value of a food reward, attractiveness of potential mate). Magnitudes, however, are not the only possible basis of reward. There is neural evidence that brain areas involved in reward processing code for frequency and temporal delay in addition to magnitude (Ellsberg, 1961). What neural computations produce these quantities is an open question, but the relevant point here is that such quantities are available in the brain. There is also a wealth of behavioral evidence demonstrating that human decision making is guided by quantities other than magnitudes. For example, in the famous paradox that bears his name, Ellsburg demonstrated that given comparable choices, people tend to prefer those with lower ambiguity concerning the nature of the underlying payoffs (Kahneman & Tversky, 2000). The framing of payoffs and their associated risks are also attributes known to shape decisions (Singh, Lewis, & Barto, 2009). Furthermore, recent theoretical work in RL has shown that that there are adaptive benefits when reward functions are themselves subject to learning via experience (Maia & McClelland, 2004). In this work, reward functions selected globally across a diversity of experiences were shown to confer local advantages in the performance of tasks compatible with those previously experienced. Decision making under uncertainty is ubiquitous in human experience, and taken together the neural and behavioral evidence and computational proposals suggest that broadening the conceptualization of reward beyond magnitude quantities may allow RL models greater flexibility in capturing human performance.

In the Base model, the reward function transformed monetary payoffs into rewards assuming that it was the net payoff that was internally rewarding to participants (c.f., Equation 3). The net payoff from a given card is certainly a salient attribute of the IGT, and best-case performance in the IGT is achieved by learning the expected values of the decks based on the trial-by-trial net payoffs. However, a sizable subset of participants perform the IGT disadvantageously therefore deviating from best-case performance. This suggests that other decision attributes may play a role in guiding choice. Further support for the role of other attributes is provided by a study in which self-reported knowledge was collected from participants as they performed the IGT. This study demonstrated that the majority of participants were aware of multiple declarative

aspects of the task, including the frequency of losses (Preston, et al., 2007). To allow for other quantities to enter into the functional mapping from deck payoffs to experienced rewards, I modeled three risk-based alternatives to net payoff magnitude as the basis of reward (Equations 7a, 7b, and 7c).

$$r(t) = -netvar(t) \tag{7a}$$

$$r(t) = -lossvar(t) \tag{7b}$$

$$r(t) = 1 - lossfreq(t) \tag{7c}$$

In these Risk-Focused models, reward was based entirely on risk quantities, either the variance in net payoffs (Equation 7a), the variance in losses (Equation 7b), or the frequency of loss occurrence (Equation 7c). In each of these Risk-Focused models, lower values of these quantities (lower variance, lower frequency) were modeled as more rewarding than higher quantities. Each of the risk quantities were computed based on the last four payoffs obtained on each deck. For example, if on the tenth trial a loss was experienced in deck C and there were no other losses from this deck in the last three trials, the loss frequency for deck C would be 25% and the reward function would yield a scalar value r(t) = 1- 0.25 = 0.75.

$$r(t) = G \cdot gain(t) + L \cdot loss(t) + \beta \cdot \left(-lossvar(t)\right) \tag{8a}$$

$$r(t) = G \cdot gain(t) + L \cdot loss(t) + \beta \cdot \left(1 - lossfreq(t)\right) \tag{8b}$$

In addition to the three Risk-Focused models, I also tested two hybrid models in which rewards were represented as a linear combination of net payoffs (as in Equation 3) and either loss variance (Equation 8a) or loss frequency (Equation 8b), with a weight parameter β determining the relative influence of the risk term on reward. As in the Risk-Focused models, the risk quantities were computed based on the last four payoffs obtained from each deck.

Methods

Data

To investigate the decision making mechanisms involved in the IGT and the ability of reinforcement learning models to capture the core phenomena associated with the task, I fit a set of RL models to behavioral data from the IGT. The behavioral data were previously collected from control participants by Preston and colleagues at the University of Iowa in a study investigating the effects of anticipatory stress on IGT decision making (Bechara, Tranel, et al., 2000). These participants were screened for

known brain damage and decision making impairments. The experimental paradigm used to collect these data was the A'B'C'D' version of the IGT first reported in (Bechara, 2007) and currently used in the commercially available version of the task used in assessment (Busemeyer & Stout, 2002; Yechiam & Busemeyer, 2005). Of the 41 participants, 9 (22%) failed to perform advantageously. In aggregate, participants chose 59 percent of their cards from advantageous decks C and D, a result consistent with standard administration of this task.

Modeling Procedures

Model fitting

The parameters for each of the models were fit to the deck selection histories independently for each of the 41 participants using maximum likelihood methods. The log-likelihood function is given in Equation 7.

$$ln\left(\mathcal{L}\left(\hat{\theta}|data_i, model_j\right)\right) = \sum_{t=1}^{T} ln\left(Prob\left(d_j(t) = d_i(t)|d_i(1{:}\,t-1)\right)\right) \qquad (7)$$

This log-likelihood function captured, for each trial *t*, the probability that model *j* selected the same deck d_j that was selected by subject *i*, given the subject's entire selection history $d_i(1{:}\,t-1)$ leading up to trial t. This log-likelihood function therefore represented the ability of a model to do one-step look-ahead for each of the T trials in the experiment. An ideal model would predict with probability equal to one the deck actually chosen by the subject on each trial, and would produce a likelihood equal to $\ln(T)$. This formulation of the likelihood is identical to methods previously reported in the literature (Busemeyer & Stout, 2002).

The parameters of each model were numerically fit to the participant data using the Nelder-Mead Simplex numerical optimization algorithm available in the Mathematica® programming language (6.0, Wolfram Research, Inc.: Champagne, IL). To minimize the chance of finding local maxima, the algorithm was run for each participant at multiple, random starting locations and the maximum value returned by these runs was used as the final result.

Model comparison and simulation

For the purposes of comparing models I adopted an approach used in (Kahneman & Tversky, 1979). Comparisons were made using the Bayesian Information Criterion (BIC) which, unlike pure goodness-of-fit criteria, penalizes a model based on the number of

free parameters. After fitting the models, I computed a relative BIC score for each model *m* compared to a null model. The null model assumed fixed selection probabilities for each deck, with the probabilities computed based on the proportion of selections from each of the decks. For example, if the total number of selections a participant made from each deck were A=10, B=15, C=35, and D=40, then the null model assumed fixed selection probabilities on each trial of 10%, 15%, 35% and 40% for deck A, B, C and D, respectively. The null model thus had three free parameters (the fourth probability can be imputed based on the other three). This model replicated the marginal selection probabilities, but did not account for any temporal patterns in the history of selections. I computed a relative BIC score for each model *m* as compared to the null model using Equation (8), in which $\mathcal{L}_m$ was the likelihood of a model computed defined within Equation (7) , and k_m and k_{null} represented the number of free parameters in model *m* and in the null model, respectively.

$$\delta BIC_m = 2 \cdot ln(\mathcal{L}_m - \mathcal{L}_{null}) - (k_m - k_{null}) \cdot ln\, T \quad (8)$$

The δ*BIC* score therefore provided a complexity-adjusted indication of the fit of a given model to the data. A positive δ*BIC* indicated that a model provided a better fit to the data than the null model, a condition which was possible only if the model was able to capture temporal aspects of participant's selection history not captured by the null model.

In addition to investigating the fit of each model relative to the null model, I tested for significant differences between pairwise δ*BIC* scores for each variant model and the Base model using the Wilcoxon Signed Rank test from which I generated standard two-sided p-values. The Wilcoxon test was appropriate for two reasons. First, the comparison of the fit of two models across the same 41 participants is a repeated measures comparison. The Wilcoxon test accounts for the repeated measures by operating on difference scores, thus providing greater statistical power. Second, a cursory inspection of the δ*BIC* scores indicated that the scores were not normally distributed, and therefore a non-parametric test such as the Wilcoxon test is preferred over parametric approaches such as a paired t-test.

In addition to using the δBIC scores to compare models, I used several other criteria that captured key behavioral phenomena in the task (c.f., Table 2.1 and Figure 2.1). A

summary of the primary selection criteria and other comparison criteria is given in Table 2.2. Several of these criteria were computed by comparing the performance of model simulations to the performance of participants. These simulations were carried out as follows. For each model, the maximum likelihood parameters for each participant were introduced into the model and each participant was simulated twenty times, with the mean values across the twenty simulations then used as the model's results for that participant. The results of these simulations for each participant were then compared to the empirical data for each participant. I compared the models using these criteria both at the aggregate level and at the level of individual participants. More qualitatively, I generated plots of selection histories for each participant based on both the empirical data and simulations using the best fit models. Comparison of these plots provided a qualitative sense for how well the models were able to capture participant decision making across the 100 trials.

Table 2.2 Model comparison criteria.

Criteria	Description of criteria
Mean δBIC (m_i , m_j) With p-value	The relative fit of model *i* compared to model *j*, using the δBIC score (model fit relative to null model) as the metric. Significance determined via Wilcoxon Signed Rank test.
%δBIC>0	The percentage of participants for which a model produces a positive δBIC score. A large value indicated that a model outperformed the null model in fitting many participants.
%Participants Best Fit	The percent of participants for which a given model was the best fitting model, when all ten models were fit to the all participants.
Selection Score Good/Bad	The mean percentage of simulated participant selections allocated to the good decks and bad decks as compared to the data.
Selection Score ABCD	The mean percentage of participants' simulated selections allocated to the each of the four decks (A, B, C, D) as compared to the data.
%EEV (model vs. data)	The mean percentage of trial selections that were from the decks with the 1st or 2nd highest experienced expected value based on the payoff history preceding a trial.
%Participant Adv. (model vs data)	The percentage of simulated participants that performed advantageously compared to the data.
Mean Run Length (model vs data)	The mean length of sequential selections that participants make from the same deck. This provides some measure of "persistence" in selections.
Mean Profit (model vs data)	The mean profit (or loss) generated by participants at the end of the experiment.

Notes. The Mean δBIC, %δBIC>0, and %ParticipantsBestFit (shaded) were the three primary model comparison criteria.

Results

Base Model Evaluation

The aggregate results of fitting the base model to each of the 41 participants are summarized in Table 2.3. As evidenced by the positive mean and median δBIC scores, the model captured temporal aspects of participant selections not captured by the null model. The model generated positive δBIC scores for 61% of the 41 participants, and 88% of the deck selections matched participants' allocations to the good (C, D) and bad (A, B) decks. Furthermore, 80% of the deck selections matched participants' allocations to each of the individual decks. On average, 65% of participant selections (and 53% of model selections) on each trial were made from the decks with the 1st or 2nd highest expected value based on experienced rewards. On average, both the model and the actual participants produced a negative profit. On average, 50% of the participants simulated by the model performed advantageously as compared to 78% in the actual data. In addition to performing more disadvantageously, simulated participants had a lower mean run length, indicating that they switched decks more frequently than actual participants.

Table 2.3 Evaluation of the base model.

Evaluation criteria	Data	Base Model
δBIC Mean		+13.6 (SD=31.6)
δBIC Median		+3.5
%δBIC>0		61%
%ParticipantsBestFit[1]		15%
Selection Score Good/Bad		88%
Selection Score ABCD		80%
%ParticipantAdv	78%	50%
%EEV	65%	53%
Mean Run Length	5.5	2.5
Mean Profit	-$437	-$260

Notes: [1]Percentage of participants best fit by model when fit of all ten models was considered.

Figure 2.4 shows the observed and simulated selection histories for a typical, advantageously performing participant. The model was able to reproduce the important behavioral phenomena including: (i) early exploration of all decks (phenomenon 1), (ii)

an early preference for disadvantageous decks A and B (phenomenon 2), (iii) a shift to a preference for advantageous decks C and D (phenomenon 3), and (iv) occasional selections from disadvantageous decks in the later trials of the task (phenomenon 4).

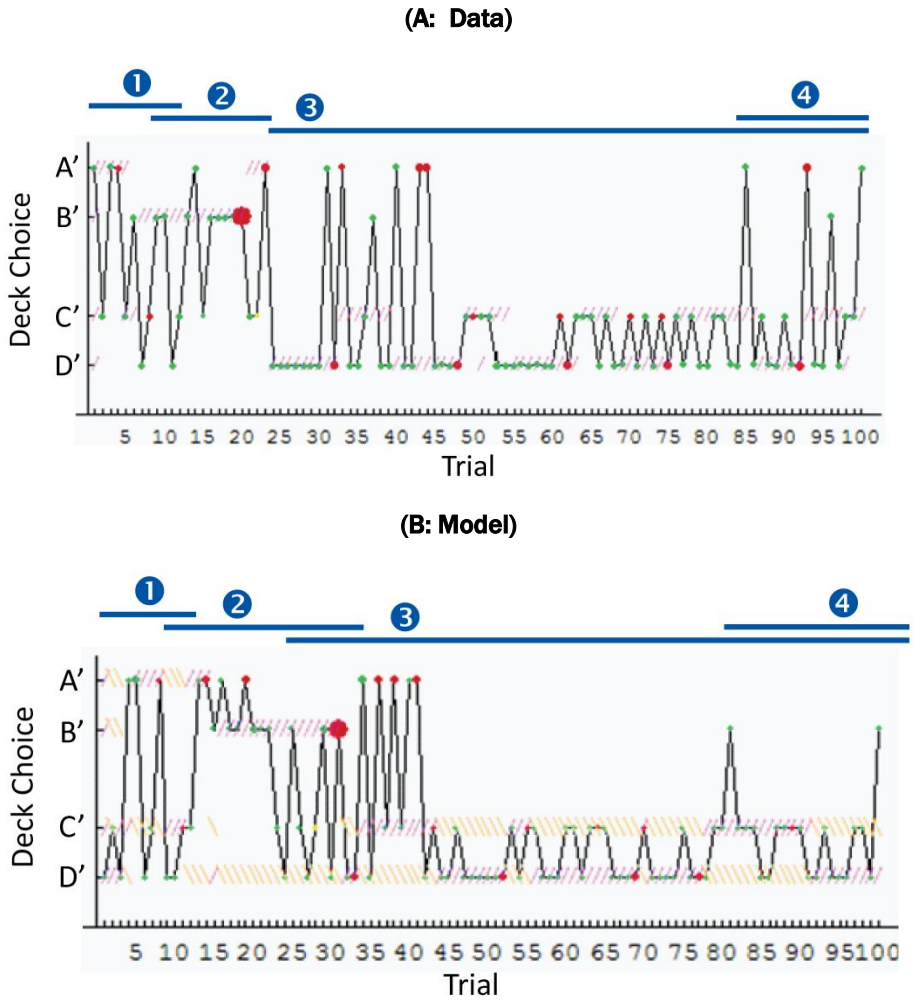

Figure 2.4. Observed and simulated selection histories. (A) Observed selection history for a typical participant who performed advantageously, and (B) the selection history for the same participant as simulated by the base model using the maximum likelihood parameters. Red dots indicate loss trials, green dots gain-only trials. The size of the dot indicates the size of the gain or loss. The pink (orange) markers indicate the deck with highest (2nd highest) expected value on a given trial based on the rewards experienced up to that trial. The numbered labels refer to the core phenomena presented in the introduction to this chapter and summarized in Table 2.1: (1) initial exploration, (2) early preference for bad decks, (3) transition to preference for good decks, (4) occasional resampling from the bad decks.

Figure 2.5 shows the observed and simulated selection paths from the advantageous decks, pooled over all 41 participants and smoothed using a 7 trial window. The model closely reproduces the overall pattern in the selection path, with early disadvantageous

selection shifting to advantageous selections after 20-30 trials, and continuing thereafter with a maximal level of 7-trial selections from the good decks reaching approximately 80%.

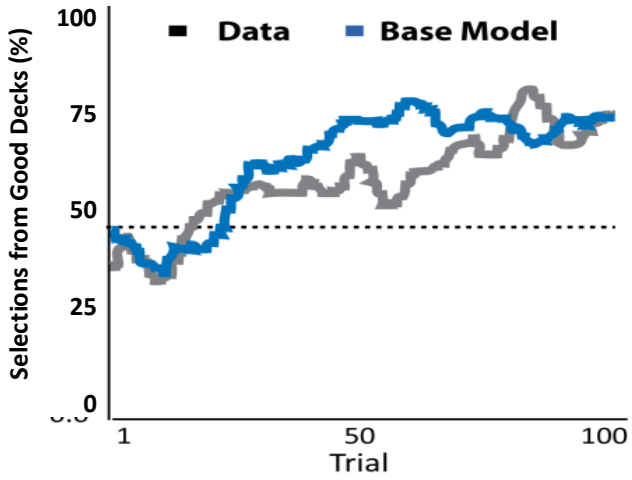

Figure 2.5. Observed (gray) and simulated (blue) mean selection path from the good decks. The figure shows the mean percent selections from the good decks) pooled over 41 participants and smoothed using a seven trial window.

Performance of the Variant Models

Simple vs. exponential averaging

I conducted the first pair wise test to determine whether the base model was improved upon by a model in the learning rate produces value estimates that are simple rather than exponential average of experienced rewards (Equation 1). The results of this model comparison are given in Table 2.4. In aggregate, the Mean δBIC for the base model was significantly higher than for the simple average model (13.6 versus 5.1. $p<0.01$). The base model fit was also superior in terms of the number of participants for which δBIC was positive (61% vs. 51%). The percentage of participants for which the model was the the best fit was 15% for both models. The relative performance of the two models was mixed across the other criteria. The Simple Average model did a better job of allocating selections to good/bad decks and to each of the four individual decks, and more closely reproduced the percent of participants who performed advantageously. It also more closely reproduced the percentage of selections from the decks with the highest expected value.

Table 2.4 Simple versus exponential averaging.

Evaluation criteria	Data	Base Model	Simple Average
δBIC Mean		+13.6 (SD=31.6)	+5.1 (SD=29.3) p<0.01
δBIC Median		3.5	0.64
%δBIC>0		61%	51%
%PartcipantsBestFit[1]		15%	15%
Selection Score Good/Bad		88%	96%
Selection Score ABCD		80%	93%
%EEV	65%	53%	65%
%ParticipantAdv	78%	50%	90%
Mean Runs	5.5	2.5	2.4
Mean Profit	-$437	-$260	-$292

Notes: [1]Percentage of participants best fit by model when fit of all ten models was considered.

Pursuit vs. softmax

In the second test, I investigated whether allowing independent storage and learning of action selection probabilities would better fit the participants. The results of this model comparison are given in Table 2.5. In terms of selection criteria, the Pursuit model significantly underperformed the Base model (Mean δBIC 4.4 vs 13.6. $p<0.01$). The pursuit model generated positive δBIC scores for only 39% of the participants. In terms of reproducing other behavioral phenomena, the Pursuit model was better along a number of the selection-based criteria, and similar in terms of profit and mean run length. The Pursuit model also fit a smaller percent of participants than the Base model (10% vs. 15%).

Table 2.5 Pursuit versus softmax choice.

Evaluation criteria	Data	Base Model	Pursuit
δBIC Mean		+13.6 (SD=31.6)	+4.4 (SD=31.9) p<0.01
δBIC Median		3.5	-2.6
%δBIC>0		61%	39%
%PartcipantsBestFit[1]		15%	10%
Selection Score Good/Bad		88%	93%
Selection Score ABCD		80%	91%
%EEV	65%	53%	58%
%ParticipantAdv	78%	50%	70%
Mean Runs	5.5	2.5	2.4
Mean Profit	-$437	-$260	-$254

Notes: [1]Percentage of participants best fit by model when fit of all ten models was considered.

Delta rule vs. reinforcement comparison

In the next test, I considered the Reinforcement Comparison model which assumed prediction error was based upon a reference reward rather than value estimates as in the delta-rule used in the Base model. The results of this model comparison are given in Table 2.6. Once again the Base model was superior in terms of fit, with a Mean δBIC significantly higher than the Reinforcement Comparison model (13.6 vs. 8.8, $p<0.01$). The Reinforcement Comparison model also fell short in terms of generating positive δBIC scores across participants (51% vs. 61%) and was best model for only 7% of the participants. On the selection and other behavioral criteria, the two models performed very similarly.

Table 2.6 Reinforcement comparison versus delta-rule learning.

Evaluation criteria	Data	Base Model	Reinforcement Comparison
δBIC Mean		+13.6 (SD=31.6)	+8.8 (SD=30.0) p<0.01
δBIC Median		3.5	0.8
%δBIC>0		61%	51%
%PartcipantsBestFit[1]		15%	7%
Selection Score Good/Bad		88%	88%
Selection Score ABCD		80%	80%
%EEV	65%	53%	52%
%ParticipantAdv	78%	50%	50%
Mean Runs	5.5	2.5	2.5
Mean Profit	-$437	-$260	-$383

Notes: [1]Percentage of participants best fit by model when fit of all ten models was considered.

The effect of value decay

Next, I investigated whether allowing the values estimates for each deck to decay might provide a better fit of the data. The results of this model comparison are given in Table 2.7. Although the Decay model had a lower mean δBIC score (12.4 vs. 13.6), this difference was not significant. The Decay model generated a positive δBIC for a smaller percentage of the participants (51% vs. 61%) and was the best model for a smaller percentage of participants (7 vs. 15%) than the Base Model. On all other criteria the two models performed similarly.

Table 2.7 Decaying versus stable values estimates.

Evaluation criteria	Data	Base Model	Decay
δBIC Mean		+13.6 (SD=31.6)	+12.4 (SD=31.6) p>0.3
δBIC Median		3.5	3.4
%δBIC>0		61%	56%
%PartcipantsBestFit[1]		15%	7%
Selection Score Good/Bad		88%	90%
Selection Score ABCD		80%	82%
%EEV	65%	53%	54%
%ParticipantAdv	78%	50%	42%
Mean Runs	5.5	2.5	2.4
Mean Profit	-$437	-$260	-$378

Notes: [1]Percentage of participants best fit by model when fit of all ten models was considered.

Risk-focused reward models

I next conducted a set of tests using the three Risk-Focused models that incorporated risk-based attributes in the definition of reward. The results of these model comparisons are given in Table 2.8. Each of these variant models performed significantly worse, in aggregate, than the Base model. These models produced negative δBIC scores indicating that they underperformed the null model which made choices based on a participants marginal deck probabilities. Interestingly, however, the Risk-Focused Loss Frequency and Risk-Focused Net Payoff Variance models provided the best fit for a small subset of participants: 7%, and 5%, respectively. This fact further motivated an investigation of model fits at the level of individual participants, which I report in a subsequent section.

Table 2.8 Risk-focused reward models.

Evaluation criteria	Data	Base Model	Net Payoff Variance	Loss Variance	Loss Frequency
δBIC Mean		+13.6 (SD=31.6)	-16.6 (SD=20.0) p<0.0001	-20.3 (SD=24.7) p<0.0001	-8.0 (SD=21.6) p<0.001
δBIC Median		3.5	-10.7	-14.2	-7.8
%δBIC>0		61%	17%	20%	22%
%PartcipantsBestFit[1]		3415	5%	0%	7%
Selection Score Good/Bad		88%	90%	90%	74%
Selection Score ABCD		80%	86%	86%	50%
%EEV	65%	53%	51%	53%	62%
%ParticipantAdv	78%	50%	46%	44%	48%
Mean Runs	5.5	2.5	2.3	2.2	18.8
Mean Profit	-$437	-$260	-$125	-$187	-$451

Notes: [1]Percentage of participants best fit by model when fit of all ten models was considered.

Risk-sensitive reward functions

Lastly, I tested the set of Risk-Sensitive models in which the reward was defined as a linear combination of net payoff and either loss variance or loss frequency. In terms of the δBIC scores, the Base model outperformed the Risk-Sensitive Loss Variance model (13.6 vs. 10.4, $p<0.01$). However, the Risk-Sensitive Loss Frequency model generated a higher mean δBIC (17.9 vs. 13.6) and a higher median δBIC (8.2 vs. 3.5) than the base model, but this difference in mean δBIC was not significant ($p>0.2$). Note that the BIC is considered the most conservative of the commonly used information criteria in penalizing models for additional parameters, and the Risk-Sensitive Loss Frequency model had one more parameter than the Base model. It is possible that the difference in model fits might be significant under a different criterion such as the Aikake Information Criterion (AIC) which imposes a lesser penalty on free parameters. The Risk-Sensitive Loss Frequency model generated approximately the same percentage of positive δBIC scores across participants (63% vs. 61%) as the Base model, but was the model for more than twice as many participants (32% vs. 15%). On all other criteria except mean profit, the Risk-Sensitive Loss Frequency model outperformed the Base model.

Table 2.9 Risk-sensitive reward models.

Evaluation criteria	Data	Base Model	Net Payoff + Loss Frequency	Net Payoff + Loss Variance
δBIC Mean		+13.6 (SD=31.6)	+17.9 (SD=32.1) p>0.2	+10.4 (SD=29.5) p<0.01
δBIC Median		3.5	8.2	0.0
%δBIC>0		61%	63%	49%
%PartcipantsBestFit[1]		15%	32%	3%
Selection Score Good/Bad		88%	91%	88%
Selection Score ABCD		80%	84%	68%
%EEV	65%	53%	64%	46%
%ParticipantAdv	78%	50%	88%	66%
Mean Runs	5.5	2.5	3.4	6.4
Mean Profit	-$437	-$260	+$198	+$107

Notes: [1]Percentage of participants best fit by model when fit of all ten models was considered.

Aggregate Model Comparison

The results for all models are summarized in Table 2.10 and are shown graphically in Figure 2.6. Across the population of participants, the Risk-Sensitive Loss Frequency model provided the best fit and the Base model yielded the second best fit. Although the difference in the Mean δBIC for the Risk-Sensitive Loss Frequency model and base model was not found to be significant, the Risk-Sensitive Loss Frequency model was the best performing model on each of the model selection criteria: it provided a higher mean δBIC score (Figure 2.6B), was the best fitting model for more than twice as many participants (Figure 2.6A) and outperformed the Null model for a slightly larger percentage of participants. This model was also better than the Base model on many of other criteria considered, for example in reproducing trial-by-trial selections and percentage of advantageous participants. Although the Simple Average model best fit the same percentage of participants as the Base model, in aggregate it produced a significantly lower mean δBIC score and outperformed the Null model for a smaller percentage of participants. The Decay model yielded a mean δBIC not significantly different than the Base model, but was the best fitting model for half as many participants. In summary, the Risk Sensitive Loss Frequency model on a range of criteria was a better model across the population than any of the other models.

Table 2.10 Summary of model performance.

Models	δBIC Mean	Participants Best Fit (%)	Participants 1st or 2nd Best Fit (%)	δBIC>0 (%)
Base	+13.6	15%	34%	61%
Simple Average	+5.1	15%	24%	51%
Decay	+12.8	7%	22%	56%
Pursuit	+4.4	10%	29%	39%
Reinforcement Comparison	+8.8	7%	7%	51%
Risk Focused: Net Payoff Variance	-16.6	5%	7%	17%
Risk Focused: Loss Variance	-20.3	0%	2%	20%
Risk Focused: Loss Frequency	-8.0	7%	7%	22%
Risk-Sensitive: Loss Variance	+10.4	3%	10%	49%
Risk-Sensitive: Loss Frequency	+17.9	32%	44%	63%

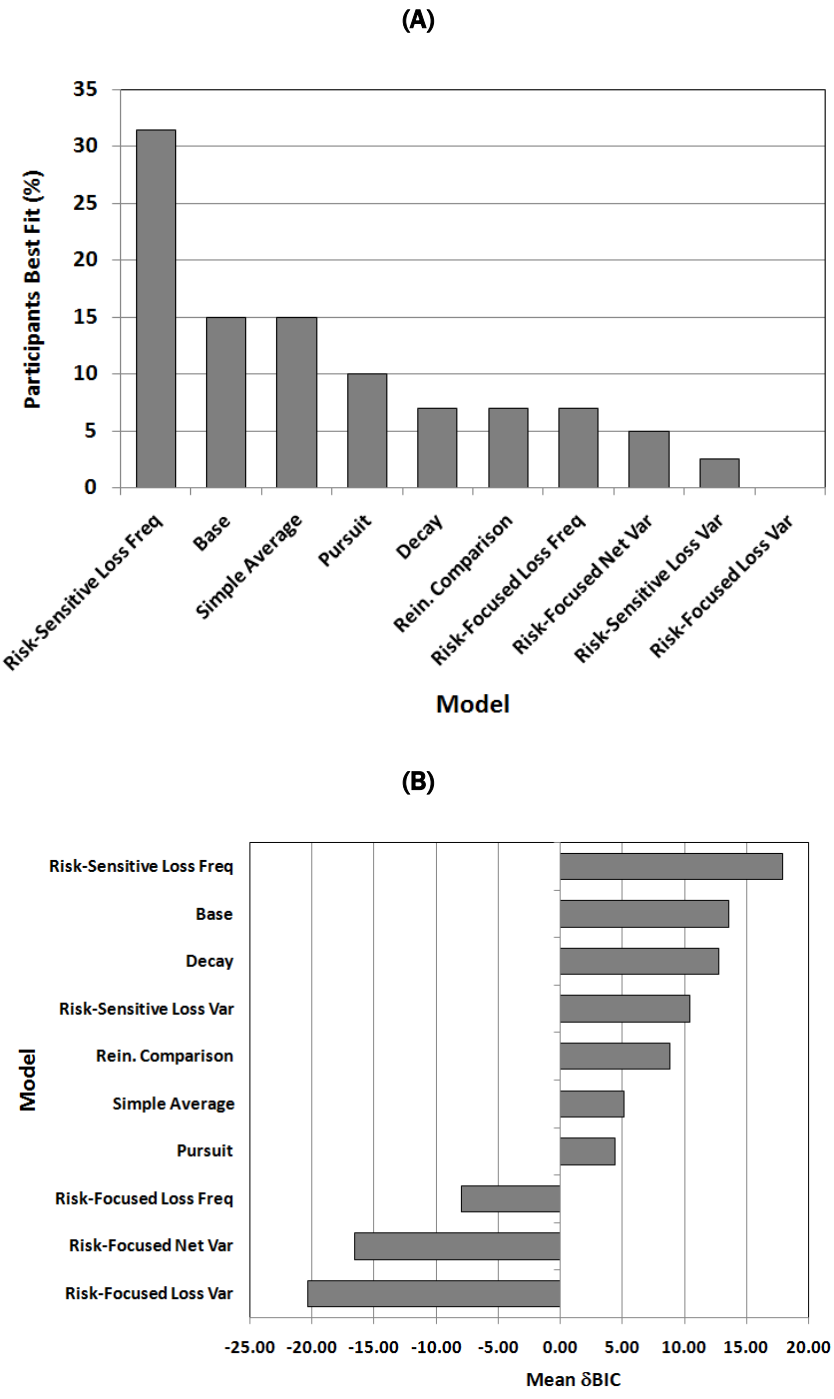

Figure 2.6 Summary of model comparison results. (A) Performance in terms of the percentage of participants best fit by each model. The Risk-Sensitive Loss Frequency model generated the best fit across the population of participants, while the other risk-based models were the worst performing models on this criterion. (B) Performance in terms of mean δBIC score across the 41 participants. On this criterion the Risk-Sensitive Loss Frequency model was also the best performing model.

Aggregate Parameter Analysis

The median maximum likelihood parameters for the two best-fitting models are shown in Table 2.11. Of interest are relative values of the parameters across models, rather than absolute parameters. For the Base model, gains were weighted more heavily than losses in the reward function. In contrast to the Base model, the weighting of gains and losses in the definition of reward in the Risk-Sensitive Loss Frequency model was approximately equal, suggesting that part of the contribution of losses was better modeled in the risk component of reward, and the β parameter in the Risk Sensitive Loss Frequency model was estimated to be 0.54 indicating that the risk and magnitude components contributed about equally to reward. With the addition of the risk component to the instantiation of reward, the estimated learning rate α for the Risk Sensitive Loss Frequency model was higher (recent payoffs matter more) than in the Base model and the choice sensitivity parameter θ was found to be lower (lesser influence of value estimates on choice). This difference in the learning rate parameters makes sense in that the occurrence of losses should weigh more heavily on value updates if losses signal the possibility of both lower expected value as well as more frequent losses, and a higher learning rate allows loss events to have greater influence on the updating of stored value estimates. In terms of the difference in the choice sensitivity parameter, one possible account is that if reward tracks both net payoffs and loss frequencies, than two attributes of choice must be tracked across the four decks and doing so is more demanding; in the face of higher demands, choice processes become more difficult and as a result are less influenced by stored value estimates and more influenced by noise. Perhaps of most interest in the estimated parameters is the finding that the addition of risk to the definition of reward resulted in sizable[1] changes in the parameter estimates of the learning rate (α), choice sensitivity (θ), and relative weighting of gains and losses (G/L). This change in parameters suggests that prior inferences using the Base model may not be robust if an important improvement to this model results in major re-parameterization for a similar population of participants.

[1] Sizable within the range of differences in parameter estimates across all of the models (not reported).

Table 2.11 Median parameter estimates for the two best-fitting models.

Model	G	L	α	θ	β
Base	0.49	0.27	0.04	0.16	-
Risk-Sensitive Loss Frequency	0.25	0.23	0.07	0.05	0.54

Individual-Level Analysis

When evaluated in terms of their overall ability to fit the population of participants, there were large differences in performance across the set of models (c.f., Figure 2.6). For example, the three Risk-Focused models failed to outperform the null model, and the Simple-Average and Pursuit models performed significantly worse than three best fitting models. However, many of the models that performed poorly across the population were the best fit for subsets of individual participants – and generated high δBIC scores for those subsets. Figure 2.7 highlights this finding by contrasting the mean fit of each model across the population (blue bars) with the mean fit of each model considering only the participants for which a model provided the best fit (green bars); the figure also shows the total percentage of participants best fit by each model (percent scores indicated along the horizontal axis). Of particular note is the contrast for the Reinforcement Comparison model. Viewed in terms of mean fit across the population, this model underperformed the Base model and the overall best fitting Risk-Sensitive Loss Frequency model. However, viewed in terms of its best-fit subset, the Reinforcement Comparison model produced the highest mean δBIC score among all the models. What this indicates is that a subset of 7% of the participants was very well fit by this model relative to the other models, and that this subset of participants was sensitive to reward relative to a learned reference rather than to absolute reward magnitudes. Similarly, the Risk-Focused Loss Frequency model and the Pursuit model generated low/negative mean δBIC scores across the population, but considered at best-fit subset level, these two models performed much better than many other models. Notably, in the Risk-Focused Loss Frequency model net payoffs were completely absent in the definition of reward, and yet this model was the best fit for 7% of the participants and produced a moderately high mean δBIC for these participants. Lastly, the Risk-Sensitive Loss Frequency model that was the best fitting model across the participants (in terms of percentage of participants best fit, and mean δBIC score) also generated a very high

mean δBIC score for the subset of participants that it fit the best, therefore providing further evidence for this model as the overall best model among those studied.

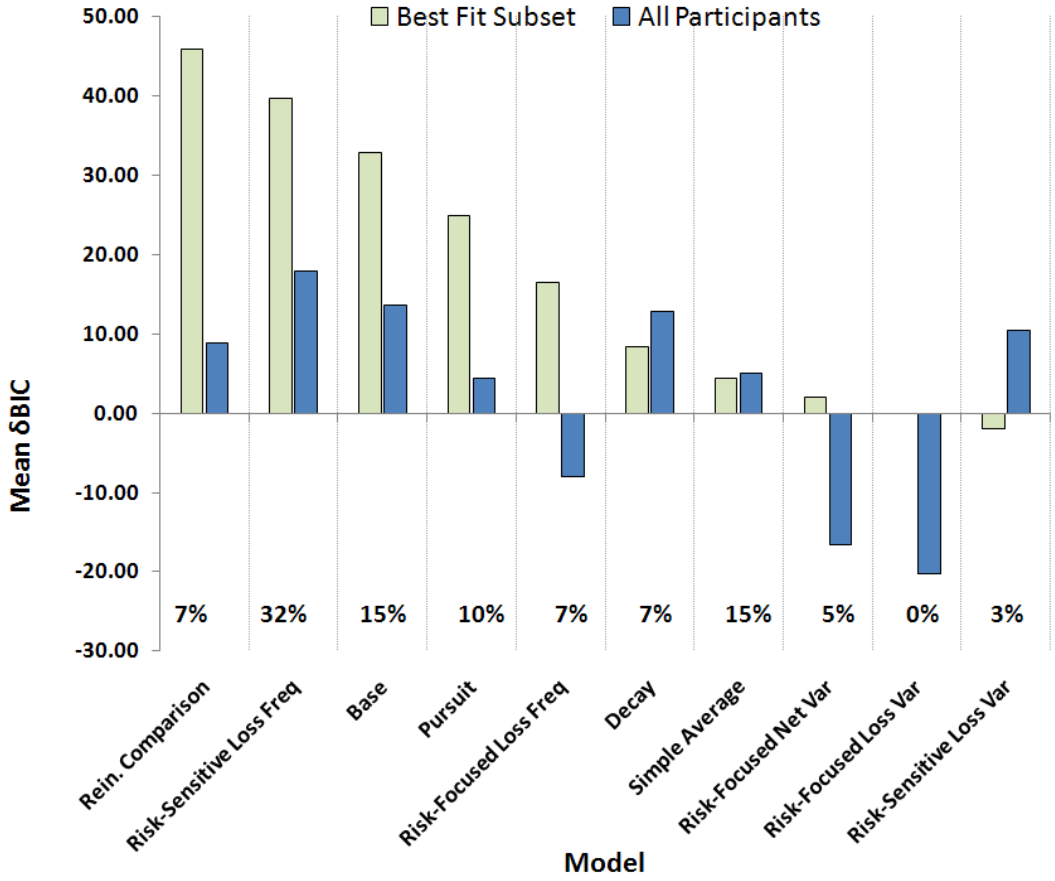

Figure 2.7. Model performance for all participants versus best fit subsets. Performance of models in fitting all participants (blue) as compared to individual participants best fit by a given model (green). The percentages listed next to the models indicate the percent of participants that were best fit by each model.

Evaluation of the models at the level of individual participants makes the findings reported in the last paragraph more clear (Table 2.12, Figure 2.8). For example, the Risk-Focused Loss Frequency model (that was among the worst overall models) produced a relatively high δBIC score in fitting Participant 1 (δBIC=40.9) suggesting that in performing the IGT what was internally rewarding for this participant was not net payoffs obtained from decks, but instead choices that generated fewer losses. Similarly, in fitting Participant 4, the Pursuit model produced the highest δBIC score of any model fit to any participant (δBIC=99.1). This suggests that for this participant, decisions may have been guided not only be learned values signaling the expected payoffs from the

four decks, but also somewhat independently by positive reinforcement of previous choices (response reinforcement). Three participants (Participants 3, 14, and 35) were best fit by the Reinforcement Comparison model suggesting that the learning of the values associated with each deck may have been shaped in part by relative comparison of payoffs obtained from each deck to a reference level of reward rather than by the absolute magnitudes of these payoffs. While the Risk-Sensitive Loss Frequency model is well supported by the data as the best model for the population, it is evident from the data in the table that no one model is able to capture all of the individual differences in decision making, despite having free parameters that were fit to individual participant data. Moreover, a large minority of participants (25-30%) were not well-fit by any of the models considered in the study (see participants indicated by the bracket in Figure 2.8).

Table 2.12 Individual-level analysis of model performance.

Participant	% Good	Best Model for Participant	Mean δBIC All Models	δBIC Best Model
15	70	Base	8.8	17.5
16	73	Base	-0.2	22
17	56	Base	40	69.1
20 (D)	46	Base	2.1	14.6
29	76	Base	19.2	30.3
41 (D)	27	Base	24.9	44.1
2	52	Simple Average	-11.1	-4.7
13	56	Simple Average	-5.2	-1
18	61	Simple Average	-3.3	2.8
22	69	Simple Average	-1.7	5.6
33	67	Simple Average	1.5	12.6
37	69	Simple Average	1.7	11.1
8 (D)	23	Decay	-14.4	-5.5
26	68	Decay	-0.3	16.7
30	59	Decay	8.5	13.8
4	76	Pursuit	52.7	99.1
5 (D)	39	Pursuit	-15.2	-4.3
9 (D)	18	Pursuit	-9.5	-0.9
23 (D)	37	Pursuit	-11.8	-2.6
3	75	Reinforcement Comparison	-0.5	36.7
14	83	Reinforcement Comparison	22.8	45.6
35	76	Reinforcement Comparison	20.1	55.4
7 (D)	32	Risk-Focused: Net Payoff Variance	-8.9	2
19	58	Risk-Focused: Net Payoff Variance	-2.7	2
1	55	Risk-Focused: Loss Frequency	-4.8	40.9
11 (D)	46	Risk-Focused: Loss Frequency	-4.2	11.0
36	52	Risk-Focused: Loss Frequency	-10.3	-2.5
12	65	Risk-Sensitive: Loss Variance	-6.1	-2
6	55	Risk-Sensitive: Loss Frequency	-28.5	-12.1
10	61	Risk-Sensitive: Loss Frequency	-5.8	46.3
21	69	Risk-Sensitive: Loss Frequency	-26	-6.2
24	56	Risk-Sensitive: Loss Frequency	-0.6	7.9
25	84	Risk-Sensitive: Loss Frequency	2.1	32.8
27 (D)	39	Risk-Sensitive: Loss Frequency	1	17.5
28	51	Risk-Sensitive: Loss Frequency	4.6	33
31	78	Risk-Sensitive: Loss Frequency	-22.1	15.9
32	76	Risk-Sensitive: Loss Frequency	109	151.4
34	55	Risk-Sensitive: Loss Frequency	2.9	13.4
38	74	Risk-Sensitive: Loss Frequency	17	38.2
39	70	Risk-Sensitive: Loss Frequency	30.5	68
40	71	Risk-Sensitive: Loss Frequency	21.6	57.8

Notes. D indicates participants who performed disadvantageously. Mean δBIC indicates how well the entire class of models was able to fit each participant. The δBIC Best Model indicates how well the best-fitting model fit each participant.

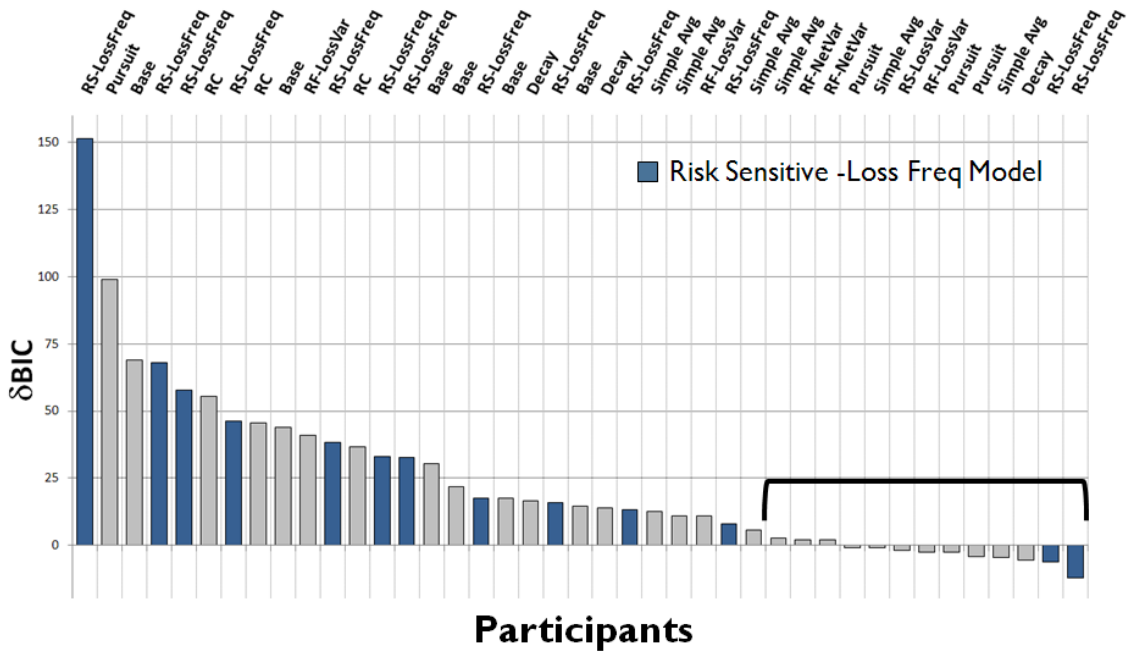

Figure 2.8. Best-fitting models for each of the 41 participants, sorted in descending order based on the δBIC criterion. Bars shaded in blue identify participants for whom the Risk-Sensitive Loss Frequency model was the best fitting model. A large minority of participants (indicated by the bracket) were not well fit by any of the models considered in the study.

Discussion

In this work I sought to formalize decision making in the IGT in terms of an explicit set of computational reinforcement learning (RL) models that were motivated by a previously demonstrated neurobiological association between frontal reward-learning areas in the brain and performance in the task. My aim was to investigate the extent to which this class of models could account for the well-documented decision phenomena associated with the task as well as for the variability in performance across individuals. Analysis in the aggregate demonstrated that the RL framework is able to reproduce, endogenously, the core features of the task: initial exploration, an early preference for the bad decks and a subsequent shift to the good decks. Comparisons between models differing in their assumptions about reward, learning, and choice mechanisms revealed that the addition of risk as an attribute of reward yielded the best fitting model across the population of participants. Specifically, when reward was augmented to include a loss frequency, the resulting model outperformed the currently accepted model of the task (as well as the other models studied) across a range of criteria, including complexity-

adjusted goodness of fit (δBIC) and the percentage of the participants best fit by the model. That this was the best performing model suggests that for many individuals, the avoidance of losses was an important component of reward in addition to the receipt of net monetary payoffs. Moreover, the fact that one of the models that considered *only* loss avoidance as reward (Risk-Sensitive Loss Frequency) provided a good fits for 7% of the participants further supports a role for loss avoidance in decision making. This is consistent with the idea of "loss aversion" that has been well-documented in the decision theory literature (2004).

Analysis at the level of individual participants, however, suggests that even the best model does not yet provide a comprehensive account of the task. No one model among those studied possessed sufficient structure to provide a good fit for a majority of participants. Furthermore, many of the models that performed very poorly across participants, turned out to be the best fitting model for sizable subsets of individuals. This was true for participants who performed advantageously as well as disadvantageously and therefore cannot be attributed to the inability of the best models to fit the poorly performing individuals.

The fact that no model in the class of models tested was able to reasonably account for individual differences for even a majority of the participants is not surprising. The standard RL framework is a model of lower-level procedural learning processes. Stimuli are processed by a procedural learning system that transforms stimuli into scalar representations of reward and then into learned values that represent the longer-term payoffs associated with the experienced stimuli and that can be used to guide choice Although the basic RL paradigm takes no stance on whether these learning and choice processes are purely implicit, explicit, or some combination, by not including higher-order cognitive functions such as problem-solving and planning, the paradigm is more consistent with lower-level procedural process commonly associated with implicit learning and less consistent with more declarative processes typically associated with higher cognition. Whether the IGT is better conceptualized as a primarily implicit task involving more of the affective or "hot" decision processes, or as a more declarative problem-solving task involving the more "cold" decision processes is the subject of considerable ongoing debate. The results of this study suggest that there are significant individual differences in the way individuals perform the IGT and that some of these

differences are not easily captured by the procedural learning processes instantiated in the RL framework. One possible explanation for these results is that higher-level cognitive processes may be at work in addition to lower-level learning processes. While its developers took an initial stance that the IGT is "cognitively impenetrable", subsequent studies have provided quite convincing evidence to the contrary. For example, Maia & McClelland (Yechiam, et al., 2005) have demonstrated that a majority of participants who performed the IGT developed declarative knowledge of the payoff structure of the task sufficient support advantageous performance. It remains to be shown whether such information is causally linked to decision behavior, but nevertheless this finding (and other related findings) suggest that computational work combining lower-level learning processes with higher-order problem solving may be worthy of attention. Alternatively, the RL models used in this study were the simplest form of model within the larger class of RL models and these were studied with functional variations considered one at a time. It is possible that by combining model variants (e.g., using simple averaging in conjunction with a loss frequency component and pursuit-based selection) greater variability might have been captured. Also, richer models certainly exist and it is possible that these models might better account for individual variability in the task. For example, it is possible that a richer representation of state that includes participants' current profit (a quantity available to participants during the game) and/or that includes informational attributes of the game (the occurrence of important events such as the first loss in a deck) might provide additional explanatory value. Hybrid models that integrate goal-oriented problem solving with the choice mechanisms of the RL framework might also an interesting direction for further research.

Lastly, the results of this study offer a note of caution for the use of the currently accepted models of the IGT to characterize clinical impairments in decision making. To date, studies using this model have relied on differences in population-averaged model parameters to infer behavioral differences between patients and healthy controls. For example, Huntington's patients and chronic substance users have been characterized as "more focused on gains and recent payoffs than healthy controls", while Parkinson's patients have been characterized as "more focused on losses and less deterministic in choice" than healthy controls (see B. D. Dunn, et al., 2006 for a review). The results of

these clinical studies rest on two important assumptions: First, that population averages provide a reasonable proxy for individual performance; and second, that the expectancy-valence model offers a reasonable (if not comprehensive) account of the decision processes that underlie behavior in the IGT. The findings reported in this study suggest that neither of these assumptions holds with great force. By adding a component of risk to the computational instantiation of reward, the efficacy of the RL-based account of the IGT was greatly improved. Critically, however, the median parameter estimates for this risk-sensitive model were very different than those found for the Base model. While it is possible that prior characterizations of patients based on model parameters might still hold in under the Risk-Sensitive Loss Frequency model, this is by no means a foregone conclusion. The base model is nested within the Risk-Sensitive model (setting the β parameter in the Risk-Sensitive model to zero yields the Base model) and that fact that median parameter estimates dramatically changed upon the introduction of risk does not bode well for the robustness of the clinical assessments that have been based on these parameters. Until a more comprehensive model of the IGT is found, it is my opinion that *model* comparisons are a more appropriate approach to clinical assessment than are *parameter* comparisons.

CHAPTER III. IDENTIFYING INDIVIDUAL DIFFERENCES

Introduction

Studies of decision making using the IGT have followed three approaches that differ in objective and method of inference (Figure 3.1). In the first approach (Figure 3.1A), the performance of a population of interest is compared to performance by a control population, and inferences are drawn based on tests of mean differences in a dependent measure, typically *%Good*. Research hypotheses most often involve proposals that the population of interest will exhibit disadvantageous performance relative to controls. Examples of such studies are numerous, and include comparisons of clinical populations to healthy controls (for example, Crone & van der Molen, 2004), comparisons across age groups (for example, O'Carroll & Papps, 2003), and comparisons of treatment to control groups under pharmacological (for example, Hinson, Whitney, Holben, & Wirick, 2006), cognitive (for example, Preston, et al., 2007), affective (for example, Brand & Altstotter-Gleich, 2008) or other experimental manipulations.

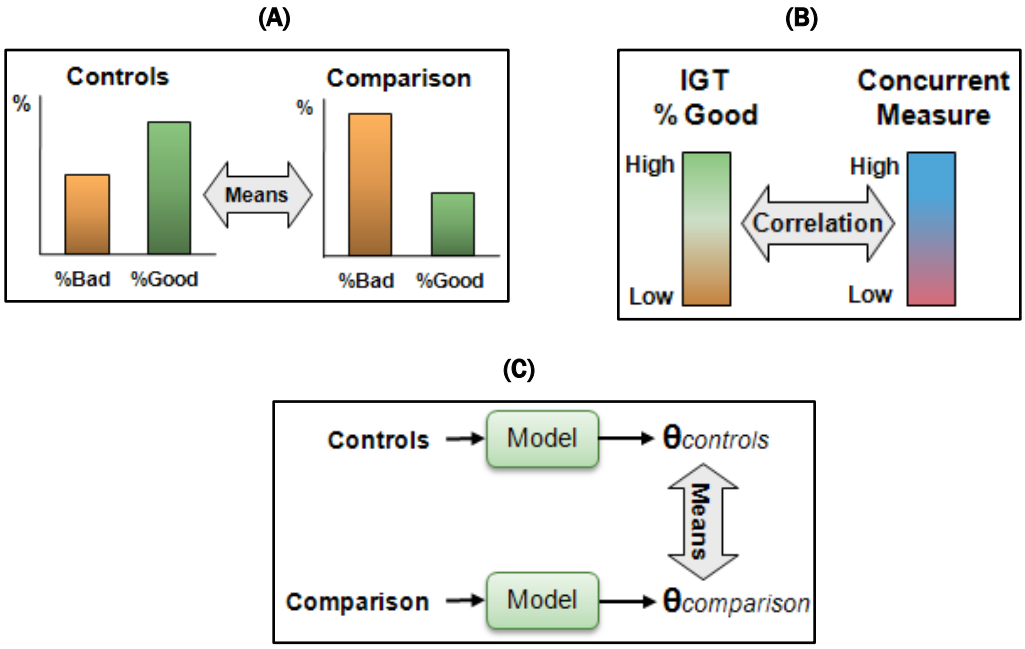

Figure 3.1 Three approaches to using the IGT to characterize decision making. (A) Clinical, developmental, and other experimental studies have compared mean differences in *%Good* to draw conclusions about decision behavior in the comparison population/condition relative to a control population/condition. (B) Correlational studies have analyzed the association between individual IGT performance in terms of *%Good* and concurrent measures such as scores from personality assessments or cognitive tasks. (C) Computational studies have fit models to individual decision data from patient and control populations and then used differences in mean parameters to characterize differences in decision behavior.

A second common approach to inference using the IGT are correlation studies that have investigated the relationship between IGT performance (typically measured using *%Good*) and concurrent measures of interest (Figure 3.1B). Examples of such studies are also numerous, and have utilized a wide range of different measures including personality traits (for example, Suhr & Tsanadis, 2006), affective states (for example, Brand, Recknor, Grabenhorst, & Bechara, 2007), and concurrent cognitive measures (for example, Yechiam, et al., 2005). A third and less common approach to inference based on the IGT has utilized computational process models to characterize differences in decision behavior across populations (Figure 3.1C). In these studies, models are fit to decision data from individual participants from two or more populations of interest. Mean (or median) model parameters are then computed for each population and parameter differences are used to characterize differences in decision behavior across the populations (Huizenga, et al., 2007).

My purpose in briefly reviewing these uses of the IGT in the study of decision making is to highlight the fact that the vast majority of these studies (i) have been based on univariate analysis of *%Good* (or *%EEV*) as a dependent measure, and/or (ii) have focused on population averages as the level of analysis. For example, clinical, developmental and experimental studies (as in Figure3.1A) have focused on population-averaged inference using univariate analyses of *%Good*. Correlational studies (as in Figure 3.1B) have narrowed the level of analysis to the individual, but are inherently univariate in terms of inference in that they attempt to predict a single dependent measure from one or more personality or behavioral measures. And computational studies (as in Figure 3.1C), in fitting a sequence of selections from the four decks, have taken a multivariate approach, but have drawn their inferences based on comparisons of population-averaged model parameters.

Given the wide ranging use of the IGT, particularly in clinical settings, one reasonable and important question is whether or not the standard univariate, population-level approach is well-justified by the data. The soundness of this approach rests on several implicit and critical assumptions. First, population-level analyses, in general, rely on an assumption that population-averaged measures are a reasonable approximation to the way most individuals perform a task. In the IGT, this implies that the choice behavior of participants is reasonably captured by a common pattern of deck selections as defined by the percentage of cards chosen from the good versus the bad decks (c.f., Figure 1.4), or the distribution of selections across the four decks (Figure 3.2A), or more stringently by a common temporal pattern of selections (Figure 3.2B). If population-level analysis is well justified, then one or more of the population-averaged patterns ought to offer a reasonable approximation to performance in the task – and in doing so provide support for inference based on the comparison of means across populations or conditions.

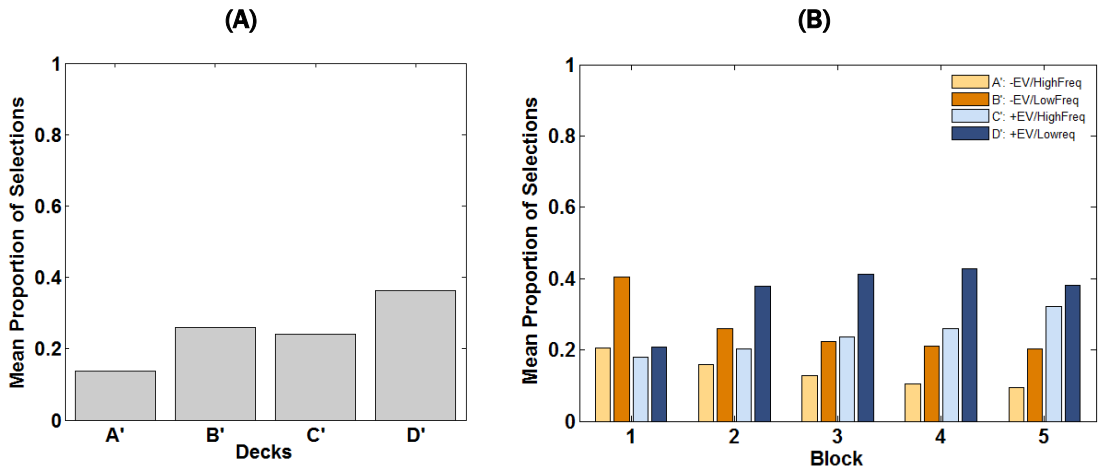

Figure 3.2 Typical IGT performance as characterized by three population-averaged patterns. (A) Mean percent of selections across the four decks and 100 trials. (B) Distribution of selections across the four decks, in five blocks of 20 trials. The values in each plot are means taken across a population of 844 participants from four independent data sets. Source: Multiple IGT data sets obtained for the purposes of this dissertation (see Methods in Chapter IV).

A second important assumption is entailed by the use of a single measure of performance, namely the assumption that individual differences are quantitative and are reasonably captured by variation along a single dimension of performance. In particular, analyses of the IGT based on *%Good* (or *%EEV*) rest on the assumption that decision behavior is guided largely by sensitivity to the expected values of the four decks and furthermore that differences in performance across participants can be explained dimensionally by differences in their sensitivity to expected value. While it is possible that IGT performance is well represented by population-averaged patterns of performance and that individual differences are well captured by quantitative differences in *%Good* (or *%EEV*), to my knowledge the strength of these assumptions has not been tested. At first look, these assumptions seem quite reasonable. Figure 3.3 shows the distribution of *%Good* across a population of participants aggregated from several independent IGT data sets. These data do not appear plainly multi-modal (Figure 3.3A), and the nearly linear pattern of variation across of the range of *%Good* (Figure 3.3B) at least suggestive that differences in performance may be dimensional in nature.

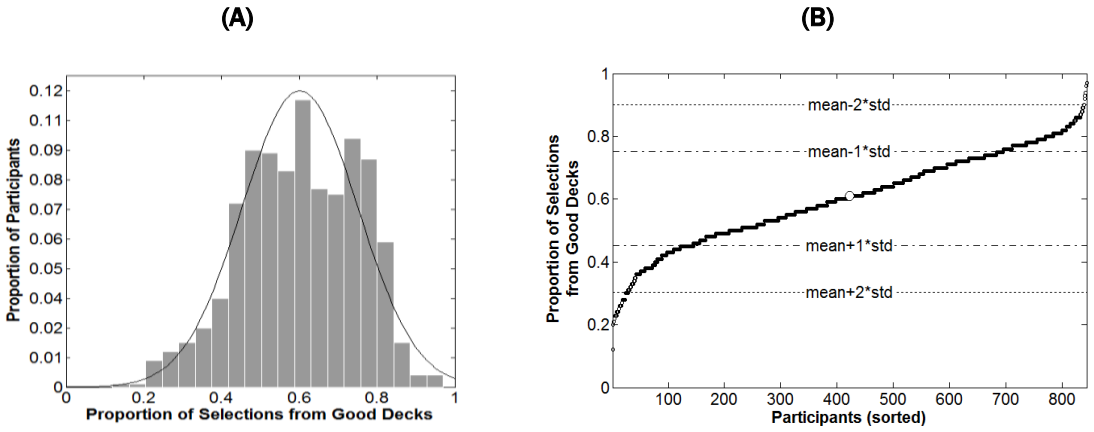

Figure 3.3. Distribution of IGT performance measured as *%Good* across a population of 844 healthy participants. (A) Histogram showing the overall shape of the distribution of *%Good*, overlaid with a Gaussian distribution fit to the data. (B) The *%Good* measure shown for each of the 844 participants, sorted from least to most advantageous. The circle marks the mean of *%Good* for the population, and the dashed and dotted lines indicate the values of %Good that are one and two standard deviations above/below the mean. Source: IGT data collected by the authors and combined for the purposes of this study (see Methods in Chapter IV).

However, there are several reasons to question the assumptions underlying standard univariate, population-level analysis of the IGT. First, there is typically a high degree of variability in the mean percentage of cards selected from each deck across participants (Figure 3.4A). Second, a large number of healthy participants typically perform the task disadvantageously. In the data set of healthy participants (N=844) used to generate the figures in this introduction, 27% of the participants selected more cards from the good decks than the bad decks (i.e., performed with *%Good* ≤ 0.5), and 14% selected more cards from the two decks with the lowest experienced value (i.e., performed with *%EEV* ≤ 0.5). Third, the results of the computational study in Chapter II revealed that a model incorporating sensitivity to risk in addition to expected value provided both a better overall fit to individual IGT decision data as well as the best fit for a large subset of the participants. Fourth, a recent study of IGT performance across age groups found a developmental trend in decision rules with guessing in young children giving way to proportional reasoning and then subsequently to value-based decision making in young adults (for example, in: Bechara, et al., 1994; Bechara, Damasio, et al., 2000; Bechara, Damasio, Tranel, & Anderson, 1998; Bechara, Tranel, et al., 2000; Preston, et al., 2007). Lastly, visual inspection of multivariate decision data at the level of individual

participants (but ignoring all temporal aspects of decision making) raises the possibility that there may be important structure in the decision data beyond what can be captured by a single measure (Figure 3.4B). While there are no well-separated clusters in the low, 3-dimensional representation of the task shown in Figure 3.4B, the data seem to be dispersed across the three diagonals suggesting the possibility of underlying structure that might be present in higher a higher-dimensional representation.

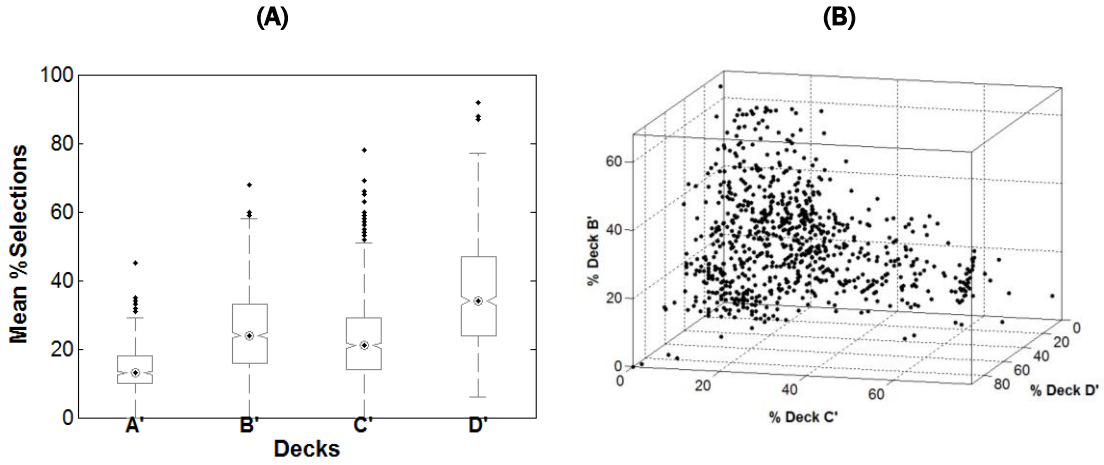

Figure 3.4 Selection variability and dispersion. (A) Mean selections by deck aggregated across a population of 844 participants from several independent data sets. The plot shows the median, 25th and 75th percentiles, range, and outliers for each deck. (B) Patterns of participant selections in multivariate space depicting the percentage of selections by each participant across decks B, C and D. Source: IGT data collected by the authors and combined for the purposes of this study (see Methods in Chapter IV).

Taken together, these facts suggest that there may be important differences in decision behavior in the IGT and I therefore sought to reexamine the task in a large sample of participants using a multivariate approach that focused on the behavior of individual participants.

Methods

I first sought to determine whether or not there exist fundamental differences in the way healthy participants perform the IGT. Based on the evidence reviewed in the Introduction, I hypothesized that decision making by participants who perform the task disadvantageously might be qualitatively different from the behavior of advantageously performing participants. I further speculated that among advantageously performing participants, there might be important differences in the temporal patterns of their

selections and/or differences in the nature of their preferences for the four decks. I tested these hypotheses using multivariate, unsupervised clustering procedures to determine whether IGT performance is well represented by a single, common decision style or is better represented by multiple decision styles. I use the term *decision style* to refer to a unique pattern of decision behavior, choosing this term because of its neutrality with respect to the ongoing debate about whether decision making in tasks like the IGT is explicit (declarative, or "cold"), implicit (affective, procedural, or "hot"), or some combination of the two.

Data Sources

The behavioral data used in this study consisted of IGT data collected from 315 participants by experimenters at the University of Iowa and reported as control subjects in the literature (Bechara, 2007). These data were also used in the development of normative data to support the version of the IGT now available as an assessment tool (Bechara, Tranel, et al., 2000). The participants were recruited from sources within and outside the university, screened for psychiatric and neurological disorders, and paid for their participation. The population of participants was 61.3% female and had mean age of 28.8 (SD of 0.79, range of 18-65) and mean years of education of 15.7 (SD =2.1, range of 11-22). These data were collected using a computerized administration of the IGT with the A'B'C'D' version of the decks. The payoff schedule of these decks is summarized in Table 1.1, described in more detail in Bechara et. al (Bechara, 2007), and is the same as the payoff schedule used in the commercially available IGT assessment tool (Hastie, Tibshirani, & Friedman, 2001). For convenience, in the remainder of this article I omit the apostrophes and refer to the four decks simply as *A*, *B*, *C*, and *D*.

Analysis Procedures

My approach to testing for the presence of multiple decision styles involved four steps. First, I generated a multivariate set of performance features for each subject. I then used these features as input to a robust clustering procedure adapted from methods developed in the literature on machine learning and data mining. Next, I used a large set of validity metrics to evaluate and select the clustering solution best-supported by the data. In the final step of the procedure, I analyzed the prototypical patterns of performance associated with the clustering solution as a basis for characterizing decision making in the task.

Feature extraction

Given the objective of investigating performance from a multivariate perspective, I extracted a set of features based on two considerations. First, I wanted the features to directly reflect the decision options available to participants as they performed the task, rather than being a set of derived measures abstracted from the task. I therefore used participants' choices among the four decks as the primary basis for measuring performance. Second, because the IGT involves learning through experience, I wanted the features to capture the temporal aspects of the task. I therefore chose to measure performance in terms of choices among the four decks across five blocks of 20 trials. Based on prior experience with IGT data, I believed that a trajectory of choices across five blocks would sufficiently capture the most important temporal aspects of task. While a more fine-grained division of choices across time might have provided additional information to a clustering procedure, I felt that such benefits would be outweighed by the additional computational complexity. Furthermore, the "four-deck by five-block" feature set provides explanatory convenience as it is consistent with the way many users of the IGT visualize their data. In summary, I chose to quantify the performance of each participant as a multivariate set of features (a 20-vector) representing the percentage of choices made from each deck, in each of five blocks of 20 trials.

The feature set therefore consisted of a [315 x 20] matrix of percentages representing the selections of four decks across five blocks by 315 participants. Because each of the 20 features were represented in percentages, they shared a common range of [0,1] and I therefore chose not to rescale the data. Furthermore, because any differences in variance across the 20 features should have been due to underlying differences in decision behavior rather than differences in unit of measure or scale, I chose not to standardize or normalize the features. For a useful discussion of the benefits and risks of data normalization and scaling prior to clustering, see Section 14.3.3 in (Duda, Hart, & Stork, 2001) and Section 10.6.1 in (Hastie, et al., 2001). As a final step in feature extraction, I looked for the presence of outliers by computing the Mahalanobis distance[2] between each participant's multidimensional performance data and the multidimensional

[2] Mahalanobis distance is a scale-invariant multivariate measure of distance (dissimilarity) that accounts for feature correlations in the data.

population mean. I then standardized these distances and removed four participants whose data were more than three standard deviations from the mean. Inspection of their performance patterns revealed that these participants either selected nearly all of their 100 cards from the two advantageous decks (suggesting that they may have had prior knowledge of the task), or they chose from one or two decks (not necessarily advantageous) and perseverated in their choice throughout the task (suggesting they may not have performed the task according to the instructions). After removal of these four outliers, the data set consisted of 311 participants.

Clustering and validity

While data clustering is well-known method in the behavioral sciences and has become an off-the-shelf tool available in most commercial statistics packages, clustering is in many ways an ill-posed problem that presents a set of challenges often obscured by its ready availability. These challenges are less serious when clustering is used for exploratory analysis. However, when identification of the underlying clusters in a data set has significant import, but "ground truth" is unknown – as is often the case in clinical assessment and medical diagnosis – these challenges loom quite large. Given the widespread and increasing use of the IGT in the study of decision making and in clinical assessment, I utilized a clustering procedure designed to address these challenges.

The objective of unsupervised clustering analysis is to use a set of features to identify subsets of samples in the data (in this study, subsets of participants) that possess a higher degree of similarity to each other than to samples that are not members of the same subset. As applied to the IGT, the goal in this study was to identify regions in 20-dimensional feature space in which there are "clumps" or clusters of participants sharing similar patterns of performance in the task (if they exist). Given no *a priori* knowledge of the statistical structure of the feature space and no ground truth on which to validate a clustering solution, clustering is inherently a difficult problem. Before presenting the procedures used in this study, I first discuss a set of well-known problems with data clustering that motivate my particular choice of methods.

In the choice of clustering methods, I sought to overcome five factors that pose challenges to the robustness and stability of clustering solutions. First, all clustering algorithms must in some way mathematically define the manner in which data samples will be judged as being similar (or dissimilar) to one another. In this study I will refer

most often to the concept of dissimilarity, but note that similarity is its inverse. It is well-established that clustering solutions can be very sensitive to the choice of dissimilarity measure, and often the choice of an appropriate dissimilarity measure is more important than the choice of the clustering algorithm itself (Dubes & Jain, 1976; Halkidi, Batistakis, & Vazirgiannis, 2002; Halkidi & Vazirgiannis, 2001; Milligan & Cooper, 1985). While dissimilarity measures are numerous, there is little theoretical guidance to assist a modeler in choosing among them and thus a dilemma: the choice of dissimilarity metric should be motivated by an understanding of the data, but the modeler typically lacks such knowledge, *a priori*. The most common solution to this dilemma is to use Euclidean distance (or for technical reasons, squared Euclidean Distance) and in fact, Euclidean distance is the default in many off-the-shelf clustering programs. A second important challenge is the fact that clustering solutions are often highly dependent on the choice of algorithm (e.g., k-means, hierarchical, etc.). Algorithms differ in their underlying assumptions and in the criteria they seek to optimize. The appropriateness of one algorithm over another depends largely on the number, shape, size, density, and separability of the clusters present in the data. So here again, a dilemma: determining whether a given algorithm is well- or ill-suited for a particular data set depends on knowledge that is often not available in advance. A third challenge is presented by the fact that many clustering algorithms (particularly those that require iterative optimization) are non-deterministic because of their dependence on how they are initialized: when run multiple times on the same data set, these algorithms can produce different solutions representing local optima or incomplete convergence to a global optimum. Another well-known problem is that clustering solutions can be highly sensitive to the number of clusters an algorithm is asked to fit, and typically the true or "natural" number of clusters is not known *a priori*. It can be the case that a solution fit to k clusters does not contain some (or any) of the same clusters found in a solution fit to $k-1$ or $k+1$ clusters. A typical approach to identifying the number of clusters (k^*) best-supported by the data is to fit solutions across a range of k and to identify the best solution based on finding the maximum (or in some cases minimum) value of a validity criterion computed for each value of k. Once again, however, the modeler is faced with a dilemma: Among validity criteria commonly used in the selection of clustering solutions, it has been found – using both artificial and real data sets in which ground truth is

known – that some criteria routinely outperform others, but no one criterion consistently selects the correct solution (Duarte, Fred, Lourenco, & Rodrigues, 2005; Sandrine Dudoit & Fridlyand, 2002, 2003; Fred & Jain, 2003, 2005, 2006; Lange, Roth, Braun, & Buhmann, 2004; Law, Topchy, & Jain, 2004; Lourenco & Fred, 2005; Topchy, Law, Jain, & Fred, 2004). A fifth challenge in data clustering is that solutions can be very sensitive to the idiosyncrasies of the particular data sample being fit, and thus do not always generalize to the population from which the sample was drawn. Small perturbations in a data sample (e.g., removal of a few samples, a slight change in the proportion of samples from different clusters, the presence of outliers, or differing numbers of outliers) can lead to significant changes in a clustering solution. This is particularly problematic in situations when data is scarce and the number of samples is small.

Clustering procedure

My approach to ensuring robustness in the clustering procedure was to exploit diversity – to rely on converging evidence from a diversity of methodological choices – rather than make single arbitrary decisions based on little or no *a priori* knowledge of the statistical structure of the IGT data. By using (i) multiple dissimilarity metrics, (ii) multiple clustering algorithms, (iii) multiple algorithm initializations, (iv) multiple validity criteria in selecting the number of clusters, and (iv) multiple realizations of the data set as produced by resampling, I addressed each of the five challenges to validity outlined in the preceding section. One difficulty in relying on a diversity of clustering methods is that clustering results must be combined in some way to determine a "consensus" solution. One approach to combination is to run all the planned variations and then to identify a final solution heuristically via examination of the resulting solutions. Another more principled approach is provided by ensemble clustering methods[3]. Ensemble clustering methods have been well-developed in the machine learning literature (Avogadri & Valentini, 2007; S. Dudoit, Fridlyand, & Speed, 2002; Luo & Liu, 2007; Masulli & Rovetta, 2003; Qiu, Wang, & Liu, 2005; Smyth & Coomans, 2007;

[3] Ensemble clustering is also referred to as consensus clustering. This method is similar in approach to bootstrap aggregation (bagging), but is more general in that it allows for aggregation across variations in method other than perturbations in the data via bootstrapping.

Yu, Wong, & Wang, 2007) and are widely employed in DNA microarray analysis and in mining gene-expression data (e.g., (Fred & Jain, 2003, 2005, 2006)), but to my knowledge are not widely used in the behavioral sciences.

In this study, I used an ensemble clustering method adapted from the evidence accumulation clustering approach of Fred and Jain (Sandrine Dudoit & Fridlyand, 2003) and the bagged clustering methods of Dudoit and Fridyland (Hartigan & Wong, 1979). A conceptual overview of the approach is shown in Figure 3.5. The main idea is to generate a large set of solutions (an ensemble) using a diversity of methods and to use these solutions as a basis for measuring the similarity between every pair of participants. Specifically, every time a clustering solution assigns two participants to the same cluster, these participants receive a "vote" as being similar to one another. Participants that are frequently assigned to the same cluster across all the solutions in the ensemble receive a large number of votes. In contrast, participants that are rarely assigned to the same cluster receive few votes. A consensus similarity matrix is produced by tallying the votes, i.e. counting the number of times each pair of participants is assigned to the same cluster. This consensus similarity matrix is then used as an input to a final stage of clustering, yielding a single *k*-cluster solution across a range of values for *k*. A set of validity criteria are used to select the number of clusters (k^*) that is best supported by the data. The efficacy of this method is rooted in diversification: participants that are frequently clustered together in spite of perturbations of the data set, differences in dissimilarity measures, differences in clustering algorithms, and variation of the number of clusters, are considered highly similar as supported by a convergence of evidence.

I used five clustering algorithms chosen because of their common usage and the diversity of their underlying assumptions. These were k-means clustering (Ng, Jordan, & Weiss, 2001), spectral clustering (Ward, 1963), and three variants of agglomerative hierarchical clustering (Duda, et al., 2001; Hastie, et al., 2001) that used the average, complete, and Ward's linkage methods (Sandrine Dudoit & Fridlyand, 2003; Fred & Jain, 2005). Each of these algorithms were applied to the data using four different measures of dissimilarity: squared Euclidean distance, city-block distance, cosine distance, and a correlation-based distance metric. Because Ward's linkage is defined only in terms of Euclidean-based distances, I did not use the other three distance metrics with this algorithm. Combining the five algorithms with four dissimilarity metrics, and taking

into account the limitations of Ward's linkage method, the procedure used a total of 17 different clustering models in generating the ensemble of solutions.

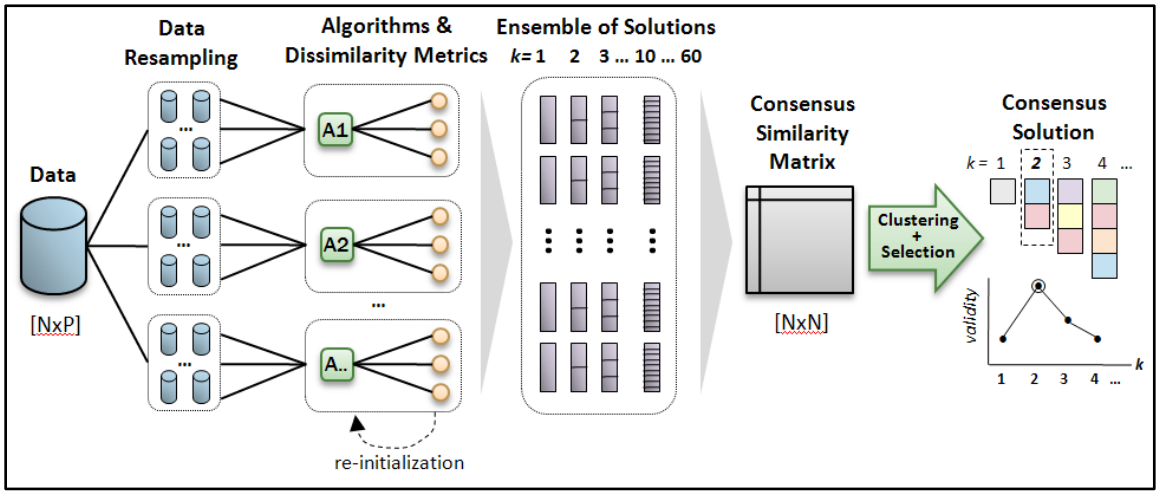

Figure 3.5 Conceptual overview of the ensemble clustering procedure. A diversity of methods is used to generate a large ensemble of clustering solutions. In the figure, this diversity includes multiple resampling of the data (bootstrapping), multiple clustering algorithms (with multiple re-initializations to avoid local optima), multiple dissimilarity measures, and fitting solutions varying in the number of clusters (k) to be fit. Other forms of diversity are also possible, for example diversity in data preprocessing and input features. An ensemble consists of solutions that assign each participant to a cluster, and the solutions vary in the number of clusters fit to the data. Participants frequently assigned to the same cluster across the solutions contained in the ensemble are considered more similar than participants that are rarely assigned to the same cluster. The ensemble therefore provides a set of votes – converging evidence – for the similarity structure in the data. A consensus similarity matrix is produced by tallying the votes, i.e. counting the number of times each pair of participants are assigned to the same cluster. A final stage of clustering is used to extract a consensus solution from the consensus similarity matrix across a range of values of k, and the number of clusters (k^*) that is best supported by the data is selected using a large set of validity criteria. In this study, I used 50 bootstrap samples (blue cylinders) of a [311x20] data set, five clustering algorithms (green squares), four dissimilarity measures (orange circles), and varied k from 2 to 15 in steps of 1 and from 15 to 60 in steps of 5, for a total of 24 variations. Method adapted from (Calinski & Harabasz, 1974).

To address the issue of generalizability of clustering solutions across data samples, I used bootstrapping to generate 50 replications of the data set of 311 participants and applied each of the 17 clustering models to each of these 50 bootstrap samples. Each bootstrap sample was generated by drawing 311 participants, with replacement, from the original data set. By sampling with replacement, bootstrapping generated replications of the data that excluded many participants and in their place included

duplicates of other participants. For the data set of 311 participants, on average a bootstrap sample will exclude approximately $(1 - 1/311)^{311}$ or 37% of the participants and therefore the bootstrapped samples represent a substantial perturbation of the data. By including solutions fit to each of these 50 bootstrap samples in the ensemble, we sought to reduce the sensitivity of the consensus solution to idiosyncratic realizations of a single data set. I note, however, that while the use of bootstrapping provides a way to enhance the internal validity of a clustering solution for a single data set, it is not a substitute for external validation using independent data sets – an endeavor pursued in Chapter IV.

Each of the 17 clustering models were fit to each of the 50 bootstrap samples, with the number of clusters (k) to be fit ranging from 1 to 15 in steps of 1, and from 15 to 60 in steps of 5, for a total of 24 levels of k. Note that k=1 is a degenerate case in which all participants were assigned to a single cluster. Although it would be extremely unlikely to find as many as 15 to 60 natural clusters in a population of 311 participants, the purpose of including solutions at very large values of k is that these solutions provide important evidence for the similarity of each pair of subjects. For example, when a subset of participants is frequently assigned to the same cluster across a set of 60-cluster solutions, this is strong evidence that these participants have similar patterns of performance[4]. Using the 17 clustering models, 50 bootstrap samples and 24 values of k, the procedure generated an ensemble consisting of 20,400 clustering solutions for the IGT data set. To guard against finding local optima with the two non-deterministic algorithms (the k-means and spectral methods), each of the 50x24=1200 runs of these algorithms was repeated 20 times with random initialization and only the best solution (the minimum squared error solution) was retained and added to the ensemble.

To extract a final solution from the consensus similarity matrix (c.f., Figure 7), I used a spectral clustering algorithm to fit k-cluster solutions for values of k ranging from 1 to 10. I then computed values for 13 validity criteria across each of these 10 extracted solutions

---

[4] Note: Including large-k solutions in the ensemble precludes the use of parametric clustering models such as Gaussian Mixtures because of the large number of parameters that need to be fit. For example, a Gaussian Mixture model with 60 clusters applied to a 20-feature data set would necessitate fitting 120 parameters in a minimal form of the model and 24,119 in a full form of the model.

and used the mean values and majority vote of these criteria to determine the number of clusters k^* best supported by the data. In general, validity criteria attempt to quantify the compactness of clusters, the separation of clusters, or some combination of the two: better solutions tend to be those with very compact clusters that are well-separated. The validity criteria used in this study were: Calinksi-Harbasz criterion (Rousseeuw, 1987), mean Silhouette width (Tibshirani, Walther, & Hastie, 2001) using Euclidean, city-block, correlation, and cosine distances, Gap-PC and Gap-Uniform statistics (J. C. Dunn, 1974), Dunn's index (Krzanowski & Lai, 1988), Krzanowski-Lai criterion (Davies & Bouldin, 1979), Davies-Bouldin criterion (Zhao, Liang, & Hu, 2006), Improved Hubert Gamma Statistic (Duda, et al., 2001), Within-Cluster Sum Squared Error criterion (Duda, et al., 2001), and the Trace Criterion (Tibshirani, et al., 2001). I abbreviate these as Ch, Sil-euc, Sil-city, Sil-corr, Sil-cos, Gap-pc, Gap-uni, Dunn, Kl, Db, Hubi, SSW, and Tr(W/T), respectively. Nine of these criteria (Ch, Sil-euc, Sil-city, Sil-corr, Sil-cos, Dunn, Kl, Db, and Hubi) select k^* as the number clusters that minimizes or maximizes the criterion. Because these criteria differ in unit of measure and whether they select based on a minimum or maximum, I linearly scaled them such that each had the range [0,1], with larger values indicating a better solution. Two of the criteria, Gap-pc and Gap-uni select k^* based a test comparing the values of the statistic at successive values of k: k^* is selected as the smallest k for which $Gap(k) \geq Gap(k+1) - s(k)$ where s is an error term (Tibshirani, et al., 2001). The two criteria SSW and Tr(W/T) are both monotonically decreasing functions of k. SSW is the total within-cluster sum-of-squares for a clustering solution and selection of k^* is accomplished by finding the inflection point ("knee") in the plot of SSW versus k. Tr(W/T) is a standardized variant of SSW in which k^* is also selected by finding a knee. To avoid the biases of determining the knee via visual inspection, I developed a regression-based approach to find the knee algorithmically.

Results

Number of Clusters

After generating the ensemble of 20,400 clustering solutions and extracting consensus solutions with the number of clusters ranging from one to ten[5], we computed and

[5] I did not expect to find more than 10 natural clusters in the data.

analyzed the validity criteria to determine whether the data better supported a single- or multiple-cluster grouping of the participants. The validity metrics for each value of k are given in Table 3.1. It is clear from these results that a single-cluster solution is rejected and that a three-cluster solution is strongly supported by the data. Of the 11 validity criteria that selected k^* at their maximum value, 10 selected the three cluster solution and the outlier (Dunn's index) selected a two-cluster solution with the three-cluster solution a very close second. The test-based Gap-pc and Gap-uni criteria selected three- and six-cluster solutions, respectively. While split in their selections, it is important to note that these are the only criteria mathematically defined for k=1, and both strongly rejected this single-cluster solution. The two "knee" criteria both selected the three-cluster solution, and therefore overall 11 of the 13 validity criteria selected the three cluster solution.

Table 3.1 Normalized validity criteria for ensemble solutions.

Validity Criteria	Selection Basis	Criteria values for number of clusters (*k*)									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Ch[a]	Max	-	0.776	1.000	0.619	0.395	0.430	0.234	0.010	0.060	0.00
Sil-euc[a]	Max	-	0.788	1.000	0.754	0.375	0.371	0.242	0.000	0.078	0.103
Sil-city[a]	Max	-	0.603	1.000	0.614	0.286	0.355	0.219	0.000	0.039	0.049
Sil-corr[a]	Max	-	0.444	1.000	0.430	0.215	0.200	0.000	0.268	0.082	0.024
Sil-cos[a]	Max	-	0.419	1.000	0.564	0.262	0.314	0.095	0.019	0.000	0.008
Dunn[a]	Max	-	1.000	0.946	0.245	0.059	0.376	0.399	0.000	0.160	0.364
Kl[a]	Max	-	0.143	0.647	0.191	0.051	1.000	0.421	0.000	0.257	-
Db[a]	Max	-	0.277	1.000	0.413	0.347	0.359	0.092	0.055	0.054	0.000
Hub[a]i	Max	-	0.483	1.000	0.511	0.421	0.165	0.198	0.000	0.102	-
Gap-pc[b]	Test	0.000	0.000	0.614	0.070	0.000	0.957	1.000	0.000	0.321	-
Gap-uni[b]	Test	0.000	0.000	0.000	0.000	0.000	0.798	1.000	0.000	0.000	-
SSW[c]	Knee	1.000	0.656	0.380	0.294	0.229	0.120	0.092	0.075	0.030	0.000
Tr(W/T)[c]	Knee	1.000	0.313	0.222	0.176	0.146	0.083	0.051	0.059	0.017	0.000

Notes: Dashes denote criteria that are not mathematically defined at a particular value of *k*. The selected value for each criterion is indicated in boldface type.
[a] Validity criteria were normalized to the range [0,1] and inverted as necessary so that the basis for selecting the number of clusters was based on their maximum value. For some criteria, selection was based on either a test comparing the value of the criterion at successive levels of k (e.g. Gap-pc and Gap-uni) or based on finding the point of maximum inflection (a "knee"). The k-cluster solution selected by each criterion is indicated in shaded cells with boldface type.
[b] The Gap criteria select k^* as the smallest *k* that satisfies the inequality Gap(k) ≥ Gap(k+1) – s(k) where s is an error term (Hastie, et al., 2001; Rencher, 2002; Tabachnick & Fidell, 2001). Values for the Gap criteria in the table are the degree to which the inequality has been satisfied i.e., the difference between Gap(k) and Gap(k+1) – s(k), with all negative values shown as zero. k^* is indicated by the first cell with a value greater than zero.
[c] SSE and Tr(W/T) are decreasing functions of *k* and therefore k^* is selected by finding the "knee" in the plot of the criterion versus *k*. To avoid the biases of visual inspection, I developedd a regression-based approach to select the knee algorithmically.

While the validity criteria strongly supported the three-cluster solution based on their "majority vote", it was also important to consider how strongly this solution was favored over the next best set of solutions. I therefore computed the means, medians and standard errors across the criteria and these values are shown in Figure 3.6. It can be seen in the figure that the mean and median of the criteria for the three-cluster solution are well separated from the other solutions. Taken together, the majority voting and the mean/median analyses strongly support the conclusion that the 311 participants in the

data set were best represented in terms of three groups that differ in their pattern of selections from the four decks across five blocks.

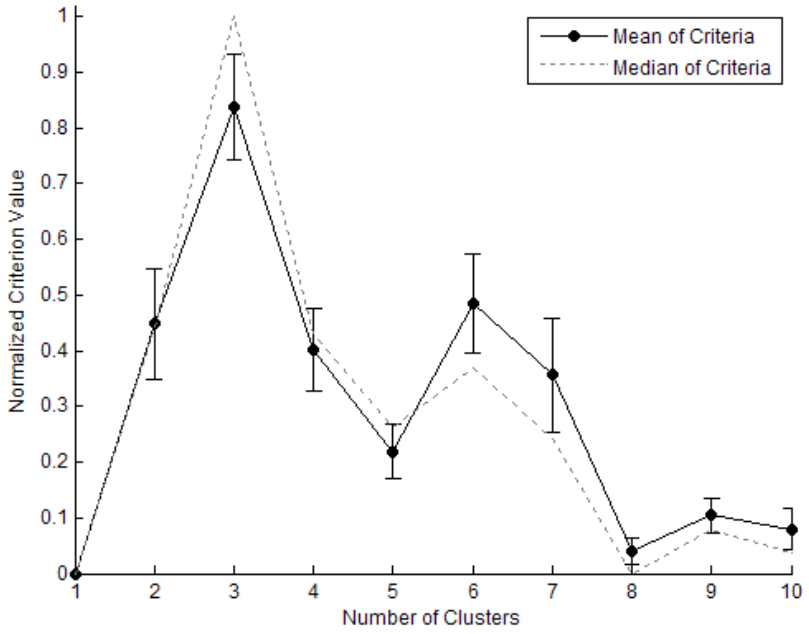

Figure 3.6 Means and medians of eleven cluster validity criteria for ensemble solutions ranging from 1 to 10 clusters. Error bars represent the standard error of the mean across the eleven criteria. Criteria were normalized to the range [0,1] and inverted as necessary so that larger values correspond to higher indicated validity. Converging evidence from this set of criteria support selection of a solution with three clusters. Not shown are two validity criteria that select k based on a "knee" in a sequence rather than based on a maximum value. These three criteria also support the three-cluster solution.

Cluster Prototypes

Having found strong support for grouping participants into three clusters, I next sought to characterize the participants in these clusters in terms of their similarities and differences in decision behavior in the task. A clustering solution simply consists of an assignment of each participant to one of *k* clusters. I used these assignments to investigate the patterns of performance for participants in these three clusters, by averaging the four-deck by five-block data separately for each cluster and then analyzing the resulting prototypical decision patterns.

The first cluster[6] prototype is shown in Figure 3.7. Of the 311 participants, 121 (39%) were assigned to this cluster by the ensemble clustering procedure. In the first block of 20 trials, these participants showed a preference for deck B, selecting approximately twice as many cards from this deck as from the other three decks from which they made approximately an equal number of selections. In subsequent blocks, these participants exhibited a strong and persistent preference for deck D, the deck that offered positive expected value and delivered low frequency losses (losses on average once every ten cards). In the final block, participants showed a slight increase in their preference for deck C. This seemed an odd change in behavior given the consistency in preference for deck D across the blocks. I therefore inspected the trial-by-trial selections for individual participants assigned to this cluster. I discovered that as a result of developing a strong early preference for deck D, many of these individuals depleted the cards[7] in deck D and as a result were forced to choose from another deck – and chose to select from deck C, the other advantageous deck. Given this fact, I concluded that after the first block, participants in this cluster did in fact show a preference for deck D that persisted for the remainder of the task. That many participants were forced to shift their selections after depleting deck D and switched to advantageous deck C underscores the interpretation that these participants were driven foremost, by a sensitivity to expected value (rather than by some other attribute unique to deck D). Hereafter, for convenience I refer to this cluster as the "EV-LowFreq" cluster.

[6] Note that clusters are typically labeled with numerical indices and the assignment and ordering of these labels is arbitrary.

[7] These data were collected using decks each consisting of 60 cards. More recent versions of the IGT include more cards so that no deck can be depleted.

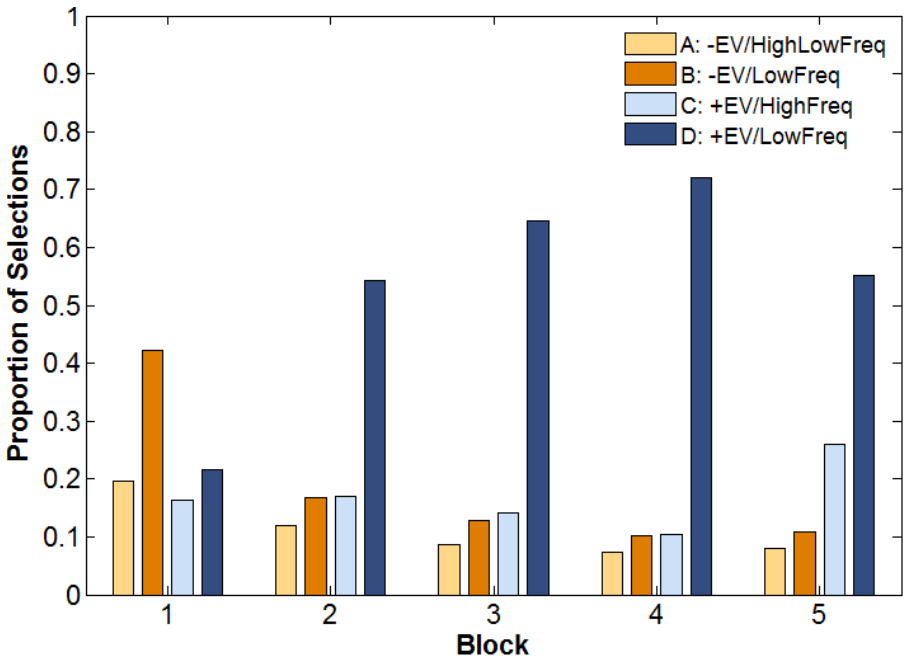

Figure 3.7 Mean pattern of performance for participants in the first cluster found in the data. This subset of participants performed advantageously. They exhibited a preference for deck B in the first block which rapidly shifted in subsequent blocks to a persistent preference for deck D, the deck that delivered positive expected value with low frequency losses. Note that the slight shift from deck D to deck C in the final block was due to the fact that many of the participants depleted the cards in deck D: forced to choose from another deck these participants chose deck C, the other advantageous deck. Of the 311 participants, 121 (39%) were assigned to this cluster.

The second cluster prototype is shown in Figure 3.8. Of the 311 participants, 76 (24%) were assigned to this cluster. In the first block of selections, these participants exhibited a pattern of selections quite similar to the participants in the EV-Low cluster, with an early preference for deck B. In the second block, their selections shifted to advantageous decks C and D, and in subsequent blocks these participants developed a clear and persistent preference for deck C, the deck that offered a positive expected value and high frequency losses (losses on average every-other card). This pattern of performance suggests that participants in this cluster were sensitive to the expected value of the decks. Hereafter, I refer to this cluster as the "EV-HighFreq" cluster and note that these participants performed the task advantageously, with none yielding a *%Good* score less than 0.57.

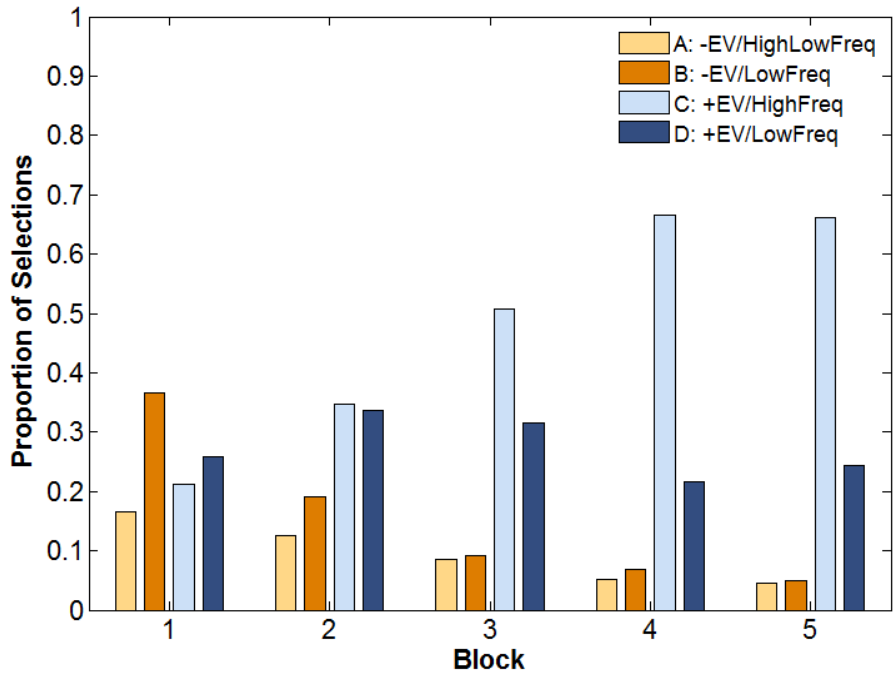

Figure 3.8 Mean pattern of performance for participants in the second cluster found in the data. This subset of participants performed advantageously. They exhibited a preference for deck B in the first block which shifted to an equal preference for decks C and D in the second block. In subsequent blocks, these participants developed an increasing preference for deck C, the deck that delivered positive expected values with high frequency losses. Of the 311 participants, 76 (24%) were assigned to this cluster.

The third cluster prototype is shown in Figure 3.9. Of the 311 participants, 114 (37%) were assigned to this cluster. In the first block of selections, these participants exhibited a similar pattern of selections as the participants in the EV-LowFreq and EV-HighFreq clusters, i.e. they developed an early preference for deck B. After the first block, these participants exhibited a clear, persistent and roughly equal preference for decks B and D, the two decks that have the common feature of delivering losses with lower frequency (on average 10% versus 50% in the decks A and C). Of the 114 participants in this in this cluster, 58 (51%) performed disadvantageously and 56 (49%) performed advantageously. Hereafter, I refer to this cluster as the "Frequency-Sensitive" cluster.

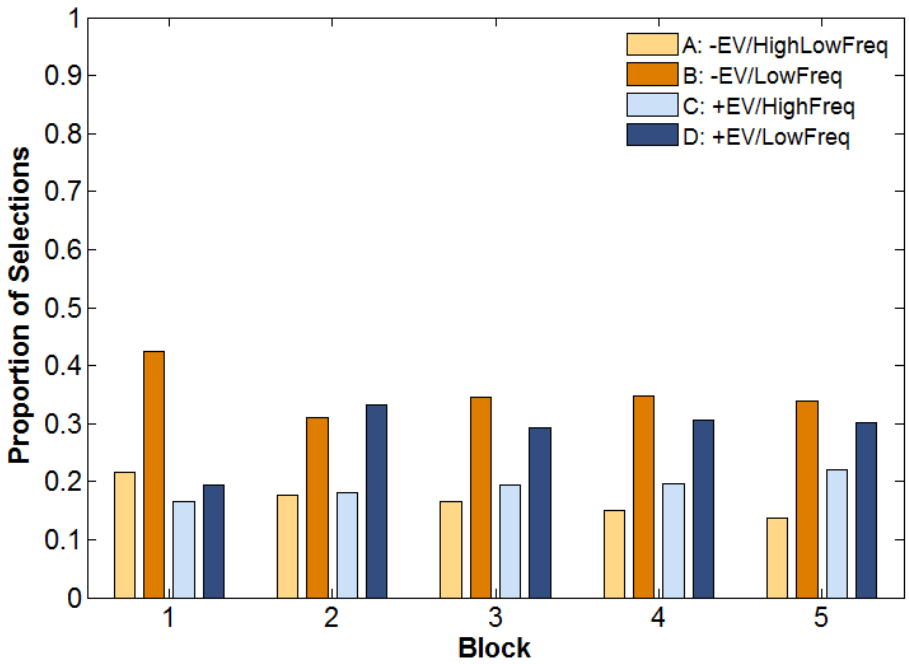

Figure 3.9 Mean pattern of performance for participants in the third cluster found in the data. These participants exhibited a combined preference for decks B and D that both delivered low frequency losses. An approximately equal number of participants in this cluster performed advantageously and disadvantageously. Of the 311 participants, 114 (37%) were assigned to this cluster.

While the results of the ensemble clustering procedure strongly supported a three-cluster solution, because of the second and smaller peak in the mean and median validity criteria in Figure 3.6, I further investigated the six-cluster solution to determine whether it represented a qualitatively different way of grouping the 311 participants, or alternatively whether it might be a variant of the three-cluster solution. By comparing the prototypical performance patterns across the six- and three-cluster solutions, I found that each of the clusters in the three-cluster solution (c.f., Figures 3.7, 3.8 and 3.9) were also present in the six cluster solution. Furthermore, I found that the additional three patterns present in the six cluster solution were variants of the three-cluster patterns of performance. Specifically, I found a variant of the EV-LowFreq cluster (c.f., Figure 3.7) in which participants' preferences for deck D developed more slowly and therefore these participants did not run out of cards in deck D. I also found a variant of the EV-High cluster in which participants' preferences for deck C developed more rapidly and as a result they depleted the cards in deck C and switched to deck D in the final block. Lastly, I found a variant of the Frequency-Sensitive cluster in which participants'

preferences for the two low-frequency decks (B and D) diminished in the later blocks, giving way to a preference for deck D as in the EV-LowFreq cluster. The fact that the six cluster solution was not qualitatively different from the three cluster solution provides additional support for the validity of the three decision patterns revealed in this study.

Discussion

Using a robust multivariate clustering method designed to minimize the dependence of the results on choice of algorithm, dissimilarity measure, initialization, and the idiosyncrasies of a single sample, I identified three subsets of participants that differed qualitatively in their performance on the IGT. These results may seem at odds with prior research on the IGT that has suggested differences in performance are reasonably characterized as quantitative differences in sensitivity to the expected values of the decks as measured by *%Good* or *%EEV*. However, by comparing and contrasting the results with prior research, I am led to interpret the findings as being complementary rather than incompatible with prevailing approaches to analyzing the IGT.

First, considering participants characterized in the literature as "advantageous", I revealed two very different decision styles. The data suggest that these two styles have in common a greater sensitivity to the longer-term expected values of the decks than to differences in the magnitude of the immediate gains obtained on each trial. The two styles, however, differed in another feature of the task, namely in the frequency with which losses occur. Within the set of the two decks with positive expected value, the participants characterized by the EV-LowFreq decision style showed a strong preference for the deck in which the losses were less frequent (deck D), while the EV-HighFreq decision style showed a persistent preference for the deck in which losses were more frequent (deck C).

The clustering results also provide new insights into the performance of healthy participants characterized in prior research as "disadvantageous" based on their performance measured in terms of *%Good*. Of the 311 participants in the data set, 58 (19%) performed disadvantageously (*%Good*<0.50). Of these participants, all of them were assigned to the Frequency-Sensitive cluster and on average these subjects drew 41% of their cards from the good decks. However, the Frequency-Sensitive cluster also consisted of an additional 56 participants who performed advantageously, drawing 54%

of their cards from the good decks. Therefore, participants in the Frequency-Sensitive decision style shared a common preference for the two low-frequency decks (B and D), but differed in the relative proportion of cards they selected from these two decks. One subset of these participants preferred deck B over deck D and as a result performed disadvantageously, while the other subset preferred deck D over deck B and as a result performed advantageously. These results suggest that rather than failing to perform the task advantageously according to the experimenter-expected objective of pursuing expected value, a large number of participants (114 of 311, or 37%) simply performed the task differently, pursuing choices that delivered losses less frequently. That there are both "advantageous" and "disadvantageous" participants in the Frequency-Sensitive cluster highlights the limitations of characterizing performance according to *%Good* (or *%EEV*) and strongly suggests that individual differences in decision behavior in the IGT are not well-characterized solely in terms of sensitivity to expected value.

CHAPTER IV. VALIDATING INDIVIDUAL DIFFERENCES

Introduction

Having revealed three decision styles using a single data set, I next investigated the external validity of the results using independent data sets. While the ensemble clustering procedure used in Chapter III provided a high degree of robustness to variability in both clustering methods as well as perturbations of the data, it is possible that the three decision styles were unique to the specific IGT data set and might not generalize to other independent IGT data. Lack of external validity could be driven by many factors including differences in test environment, task instructions, participant incentives and motivations, participant demographics, and language and culture, to name only a few of the possible sources of variability.

I sought to test the external validity of the results in two ways, via replication and via prediction. If the three decision styles found in Chapter III were robust, then applying the same ensemble clustering methods to independent IGT data sets should replicate the findings of Chapter III, producing: (i) a solution with the three clusters and (ii) a similar set of prototypical patterns for each of the three clusters. Satisfying these requirements would provide strong support for the external validity of the three decision styles. I also sought to challenge external validity in a different way, by investigating how accurately a clustering solution fit to one data set could predict the decision styles of participants from the other data sets.

Methods: Replication

Data Sources

For this study I used four independently collected data sets (Ind1-4) in addition to the data set of 311 participants used in Chapter III (Base). These data were provided by five different primary investigators, in four different laboratories, at three universities, in two countries spanning a period of approximately five years. In each of these data collections, the IGT was administered using the same payoff schedule (*A'B'C'D'*) and the same computerized administration software. The participants in each of these data sets

were used as the control population in the experiments for which they were collected. All participants in these studies were adults, were screened for neurological and psychiatric disorders, and were given monetary payments or lottery tickets to win prizes based on their performance in the task; an exception are the participants in the Base data set who were paid on an hourly basis rather than based on their performance. Summary information for each data set is given in Table 4.1. Across these five populations of healthy adults, participants vary in age (lowest mean age 18.73, highest mean age 32.28) and relative proportion of males and females (lowest percentage of females 0.32, highest percentage 0.78). While participants in each of the data sets in aggregate performed advantageously, a one-way analysis of variance revealed a significant difference across the five sets, $F(2,884)=9.848$, $p<0.001$. Post-hoc, Bonferroni-corrected comparisons ($\alpha=0.05$) indicated that the higher mean values of *%Good* in the Base and Ind3 data sets were significant relative to the mean values of *%Good* in the Ind1 and Ind2 data sets.

Table 4.1 Data sets used for external validation.

Data Set	N	Mean (std) Age	Percent Female	Mean (std) %Good	Mean (std) %EEV	Country, Language, University
Base	311	28.82 (9.84)	0.61	0.63 (0.15)[a]	0.70 (0.16)	U.S.A, English, Univ. Iowa
Ind1	352	18.73 (1.07)	0.78	0.57 (0.15)	0.65 (0.15)	U.S.A., English, Univ. Iowa
Ind2	73	31.48 (8.09)	0.32	0.57 (0.15)	0.64 (0.14)	Rep. Korea, English, Catholic Univ.
Ind3	110	21.73 (6.16)	0.56	0.64 (0.15)	0.69 (0.17)	U.S.A., English, Univ. Michigan
Ind4	39	32.28 (10.69)	0.50	0.59 (0.15)	0.66 (0.14)	U.S.A., English, U. Iowa

Notes. [a] The presence of nearly identical standard deviations is chance result and not due to miscalculations. These values differ, but identical when rounded to two digits.

Analysis Procedures

To investigate whether or not there was evidence of the three decision styles in the four independent data sets, I performed the identical ensemble clustering procedure that was used in Chapter III on each of these data sets. As in Chapter III, prior to clustering I computed Mahalanobis distances to check for the presence of outliers using a z-score of 3.0 as a cutoff. This procedure led to the removal of 4, 5, 1, 3, and 1 participant in the Base, Ind1, Ind2, Ind3, and Ind4 data sets, respectively. I then ran the ensemble

clustering procedure, identified the number of clusters selected in each data set by majority vote of the 13 validity criteria (c.f., Table 3.1) and a confirmatory examination of a plot of these criteria (c.f., Figure 3.6). I also plotted and compared the prototypical patterns of performance for each cluster across the five data sets, to determine whether or not these were qualitatively similar or different patterns.

Results: Replication

The results of applying the ensemble clustering procedure to each data set are shown in Table 4.2 and Figure 4.1. For each of the four independent data sets, the converging evidence provided by the validity criteria selected a three-cluster solution. For each of the data sets, Table 4.2 and Figure 4.1 show the mean values of 11 normalized validity criterion for clustering solutions ranging from 1 to 10 clusters. Consistent with the results from Chapter III, a three-cluster solution provides the best fit to each of the data sets. The validity criteria not included in the mean values shown in the table were excluded because they select the number of clusters based on finding a knee (SSW, Trace(W/T) and therefore are not commensurate with the others for the purposes of computing means. These two criteria selected either the three- or four-cluster solution in each of the data sets, and consideration of these criteria in the analysis did not change the majority vote for a three cluster solution in any of the data sets.

Table 4.2 Results of ensemble clustering across IGT data sets.

Data Set	Selected k^*	Mean value of validity criteria[a] for number of clusters (k)									
		1[b]	2	3	4	5	6	7	8	9	10
Base	3	0.00	0.45	**0.85**	0.40	0.22	0.48	0.36	0.04	0.11	0.08
Ind1	3	0.00	0.31	**0.82**	0.37	0.43	0.10	0.26	0.18	0.24	0.03
Ind2	3	0.40	0.59	**0.87**	0.76	0.55	0.26	0.20	0.35	0.12	0.11
Ind3	3	0.00	0.73	**0.80**	0.68	0.62	0.48	0.20	0.17	0.29	0.10
Ind4	3	0.25	0.48	**0.83**	0.66	0.53	0.26	0.36	0.27	0.23	0.07

Notes. Selected value of k for each data set is shown in boldface type. [a] Mean of eleven validity criteria that select k^* at the maximum value of the criterion. [b] Mean value at k=1 includes only the two criteria (Gap-pc, Gap-uni) that are mathematically defined at a one cluster solution.

Figure 4.1 shows a plot of the mean values of the 11 criteria for each of data sets, highlighting the support for the three-cluster solution. The lack of a clear peak in the validity criteria for the six-cluster solution that received some support from the Base data

set (*c.f.*, Table 3.1 and Figure 3.6) suggests that this solution was an artifact of that data set –likely due to the depletion of cards in decks C and D by some of the participants. Furthermore, there is little agreement across the data sets for solutions at the larger values of *k*.

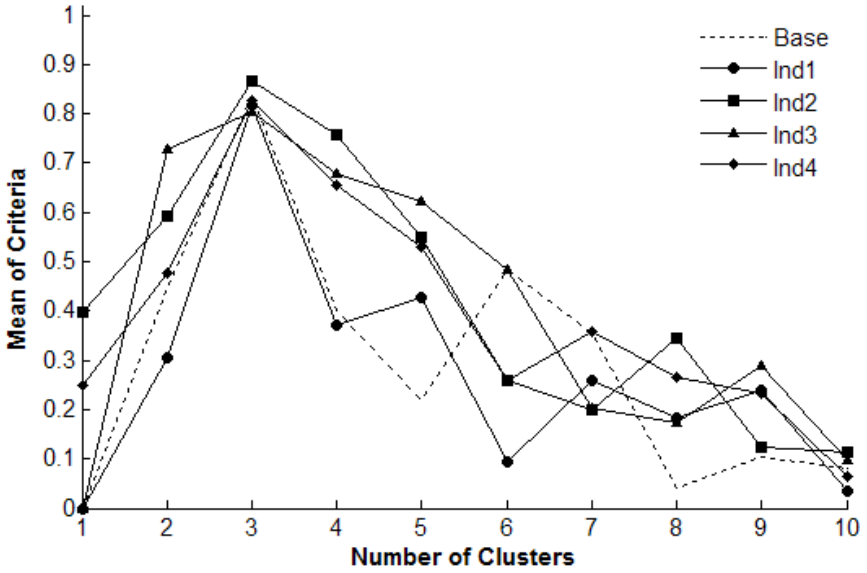

Figure 4.1 Mean values of eleven validity criteria for solutions consisting of 1 to 10 clusters. The results for the Base data set from Chapter III are shown by the dotted line, for reference. Converging evidence from the validity criteria support a three-cluster solution for each of the five data sets.

Having identified a three-cluster solution for each data set, I investigated the prototypical patterns of performance associated with each of the three clusters, for each data set. To facilitate comparison, I present these prototype patterns as mean selections from the four decks in total, rather than across five blocks as presented in Chapter III. Comparison of prototype patterns required that these patterns be matched across the data sets to identify which patterns in each of the independent data sets most closely matched the EV-LowFreq, EV-HighFreq and Risk Sensitive patterns found in the Base Data. Because I found that the patterns were very similar across the data sets, I was able to do this matching by visual inspection[8]. I first consider the results of comparing the prototypical patterns most closely matched to the EV-LowFreq pattern (Figure 4.2). It

[8] Note, however, that the matching process could be done algorithmically using exact (checking all permutations) or approximate methods (greedy search), depending on the number of clusters to be matched.

can be seen in the figure that the ensemble clustering procedure found the EV-LowFreq decision style in each of the data sets. While there are small quantitative differences in the patterns across the five data sets, it is clear that there were participants in each data set that exhibited a strong preference for deck D that delivered positive expected value with low frequency losses. In addition to the clear preference for deck D shown in the figure, the ordering of preferences over the four decks is also perfectly replicated in each of the five data sets ($D \succcurlyeq B \succ C \succ A$).

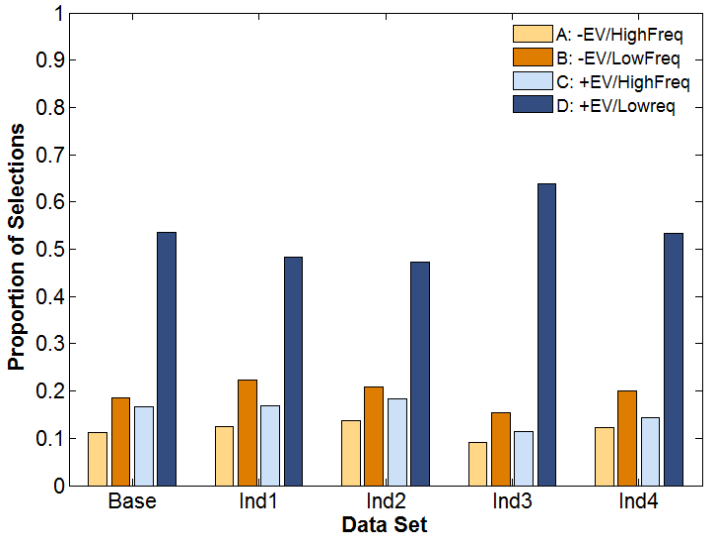

Figure 4.2 Prototypical patterns (mean proportion of selections) most closely matched to the EV-LowFreq cluster identified in the base data set. While the prototypes show small quantitative differences across the five data sets, they represent the same qualitative pattern of performance: a clear preference for deck D that delivered positive expected value and low frequency losses.

I next consider the results comparing the best-matched patterns for the EV-HighFreq cluster (Figure 4.3). It is evident from this comparison that participants exhibiting the EV-LowFreq decision style were present in each of the data sets. As in the comparison of the EV-LowFreq cluster, there are small quantitative differences, but the prototypical pattern is remarkably similar across the data sets and the ordering over preferences is identical ($C \succcurlyeq D \succ B \succ A$). These participants exhibited a clear preference for the deck that delivered positive expected value with high frequency losses, selecting 39-48% of their cards from this deck.

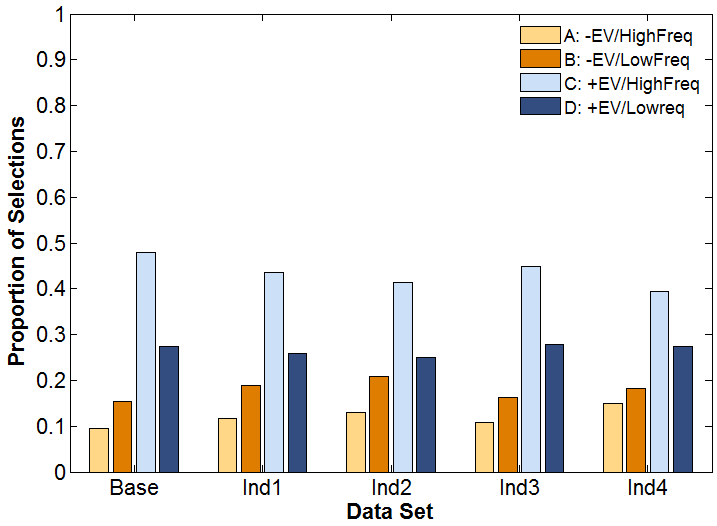

Figure 4.3 Prototypical patterns (mean proportion of selections) most closely matched to the EV-HighFreq cluster identified in the base data set. While the prototypes show small quantitative differences across the five data sets, they represent the same qualitative pattern of performance: a clear preference for deck C that delivered positive expected value and high frequency losses.

Lastly, I checked for replication of the Frequency-Sensitive cluster across data sets (Figure 4.4). Again, the results show that the clustering results found in the Base data set generalized well to the other data sets. In all data sets, participants assigned to this cluster showed a preference for the two decks that have in common the fact that they delivered lower frequency losses. There are quantitative differences in the patterns, most notable for the Ind3 data set in which the preference for decks B and D are more balanced than in the other data sets in which deck B is preferred over deck D. As in the other two clusters, the ordering of preferences over the four decks was identical across the data sets ($B \succ D \succ C \succ A$).

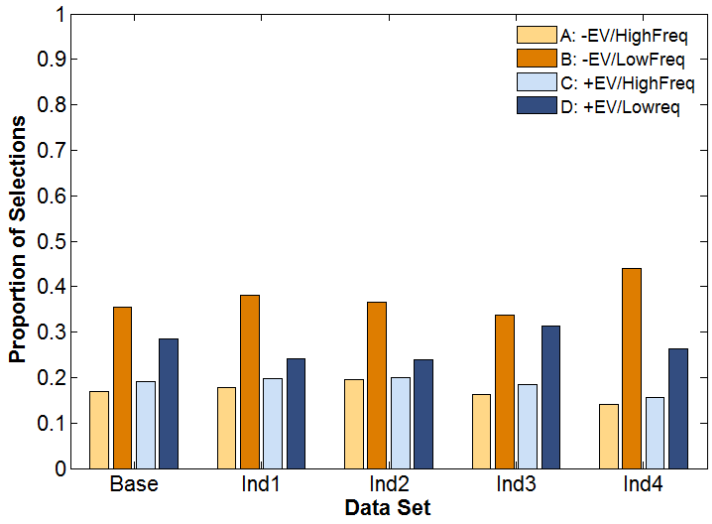

Figure 4.4 Prototypical patterns (mean proportion of selections) most closely matched to the Frequency-Sensitive cluster identified in the base data set. While the prototypes exhibit small quantitative differences across the five data sets, they represent the same qualitative pattern of performance, namely a preference for decks B and D that delivers deliver low frequency losses relative to the other two decks.

Having found the same number of clusters in each of the data sets, and determined that they represent the same qualitative decision styles found in the Base data set, I was interested in whether or not there were large differences in the relative proportions of participants assigned to each decision style across the five data sets. A high similarity in the distribution of participants across decision styles would provide additional evidence to support the generalizability of the results. To investigate this, I identified the proportion of participants found in each of the three clusters in the Base data, and these are summarized in Table 4.3. While there are quantitative differences, overall I found a very similar distribution of participants across the three decision styles. With the exception of the Ind4 data set, I found a consistent minority of participants in the EV-HighFreq cluster. This decision style seems to represent, on average, approximately 25% of the total population of 885 participants that I studied. The EV-LowFreq and Frequency-Sensitive decision styles seem to be present in approximately equal proportions, with participants assigned to the Frequency-Sensitive cluster being the majority in four of the five data sets. To test whether the distributions of participants across clusters in any of the four independent data sets differed significantly from the proportions found in the Base data set, I conducted χ^2 goodness-of-fit tests using the frequencies from the Base data set as expected frequencies against which to test the

observed frequencies for each of the other four data sets. The results of these tests are given in the last column Table 4.3 and indicate that the null hypothesis cannot be rejected for any of the data sets, suggesting that the distribution of participants by decision style generalized across independent data sets. Finding similar distributions across independent data sets is certainly not necessary evidence for the external validity of the clustering results. The data sets differ in mean age and in gender ratio (and possibly other factors), so any associations between these factors and decision style could certainly impact these distributions. However, the fact that I did not find large differences in the size of the three clusters is certainly comforting.

Table 4.3 Distribution of participants across decision styles.

Data Set	N	EV-LowFreq	EV-HighFreq	Frequency-Sensitive	$\chi^2(0.01, 2)$
Base	311	121 (39%)	76 (24%)	114 (37%)	-
Ind1	352	138 (39%)	69 (20%)	145 (41%)	5.35 (p=0.079)
Ind2	73	23 (32%)	20 (27%)	30 (41%)	1.68 (p=0.431)
Ind3	110	38 (35%)	27 (25%)	45 (41%)	1.08 (p=0.582)
Ind4	39	11 (28%)	16 (41%)	12 (31%)	5.91 (p=0.052)
Total (Mean %)	885	331 (37%)	208 (24%)	346 (39%)	-

Methods: Prediction

Data Sources

The data set used for prediction was the same set used in the replication analyses reported in the previous section.

Analysis Procedures

I sought to provide additional support for the external validity of the findings by assessing the degree to which the clustering solutions identified using one sample of data were able to generalize by predicting the cluster membership of participants in other independent data sets. For this prediction analysis I used discriminant classification methods (Rencher, 2002; Tabachnick & Fidell, 2001). The basic idea of discriminant classification is to assign a set of unknown test observations (in this study,

participants) to a particular group (in this study, cluster/decision style) based on a model obtained from a set of known training observations. In discriminant classification, the model is a based on a set of discriminant functions that represent boundaries of maximum separation between groups of observations in multivariate feature space. In this study, as in Chapter III, the feature space was represented by 20-dimensional vectors that were the percentages of selections in the four decks across five blocks. Each of our data sets therefore consisted of these 20-dimensional vectors, one for each of the participants in the data set. As in Chapter III, a clustering solution was simply an assignment of each of the participants to a particular cluster (decision style). To do prediction analysis, we fit discriminant models using the cluster assignments and feature vectors from a set of training data (one of the five data sets or a combination of the data sets), and I then used this trained model to predict the cluster (decision style) of participants from test data sets.

One of the most commonly used discriminant classifiers is based on linear discriminant functions. Linear discriminant classifiers, however, can produce poor solutions when there is heterogeneity in the covariance matrices across the groups being classified (Hastie, et al., 2001). I tested for heterogeneity in the covariance matrices in the Base data set using Box's M-test (χ^2 approximation) which rejected the null hypothesis of homogeneity ($M=146.66$, $\chi^2_{(.05,420)}=133.03$, $p<0.0001$). I therefore chose to perform quadratic discriminant classification (QDC), a common approach used when the covariance matrices are heterogeneous across groups[9]. I also sought to confirm the QDC results using a k-nearest-neighbor classifier (KNN) which is a non-parametric approach to classification.

My objective in this prediction analysis was not to identify a maximally accurate classification model by testing a broad array of different models, but rather to assess how well a reasonable model trained on one set of IGT data might generalize in predicting another set of independent data. Given this objective, I was therefore interested in comparing actual classification accuracy ("between-data" accuracy; accuracy in

[9] In quadratic discriminant classification, separate covariance matrices are used for each group rather than a single pooled covariance matrix as used in the linear discriminant approach. I used a parsimonious form of the quadratic model, one which accounts for differences in variances across groups, but ignores covariances.

predicting test data) to apparent classification accuracy ('within-data" accuracy; accuracy in predicting the training data itself[10]) as well as to chance. In all cases, a prediction was considered correct when a model assigned a participant to the same cluster as was assigned by the ensemble clustering procedure. I computed apparent accuracy using leave-one-out cross-validation (B. D. Dunn, et al., 2006): I trained a classifier using training data with one participant left out and I then determined whether this participant was assigned to the same cluster that was assigned by the ensemble clustering procedure; I repeated this process for each participant in the training data, and computed apparent accuracy as the mean percent accuracy across all participants.

I conducted two types of prediction tests. First, given that Chapter III focused entirely on the Base data set, for continuity across studies I thought it would be useful to understand how well the three clusters found in the Base data generalized to independently collected data. I therefore trained a classifier model on the features and three-cluster solution for the Base data set, and tested its accuracy in predicting the decision style for participants in one of the other independent data sets. I performed this type of pairwise prediction test using the Base data set and each of the four independent data sets (c.f., Table 4.1). I considered these strong tests of generalizability, as any unique differences between the independent data sets and the Base data set would be highlighted in the pair wise prediction accuracies, thus offering the possibility of revealing a data set for which the clustering results from Chapter III did not generalize. Although I identified and characterized a three-cluster solution in Chapter III using the Base data set, there was no *a priori* reason to believe that this data set should serve as "the" solution to use in evaluating the generalizability of the results. I therefore performed a second set of prediction tests that I believed would provide a better measure of the generalizability. I combined the five data sets into a single data set that consisted of 885 participants and then used leave-one-out cross-validation to measure classification accuracy. I trained a classifier using data from 884 participants, tested the classifier using the one participant excluded from the data, and repeated this process 884 times to yielding a mean classification rate across the 885 tests.

---

[10] It is often the case that a classifier is less than 100% accurate when tested on the same data on which it was trained.

Results: Prediction

Table 4.4 gives the results of testing how well the three clusters found in the Base data serve as predictors of the decision styles of the participants in each of the four other independent data sets. I remind the reader that the objective was not to identify a maximally accurate classifier for predicting IGT performance, but rather to use a reasonably accurate classifier to test how well the clustering solution from Chapter III generalized to predicting other data. As a baseline, the within-data set prediction accuracies for the QDC and KNN classifiers were 93.6% and 95.8%, respectively. Chance for these predictions was 33%. We therefore concluded that both of these classifiers met the requirement for reasonable accuracy. The ability of these classifiers to predict decision style in the four independent data sets is given in Table 4.4. Overall, the classifiers demonstrated a high level of accuracy, with mean accuracies of 84.9% (8.7 percentage points lower than baseline) and 83.2% (12.8 percentage points to baseline) for the QDC and KNN models, respectively (chance was 33%). Put differently, what these results indicate is that as a basis for predicting the decision style of IGT participants, there is only about a 10% loss in accuracy when predicting across independent data sets (differing in size, mean age, gender ratio, experimental methods, and culture and language) as compared to prediction within a single data set. I take this as solid evidence for the external validity of the results of Chapter III.

Table 4.4 Pairwise tests of prediction accuracy.

Pairwise test	N	QDC accuracy	KNN[b] accuracy
Base / Base[a]	311	93.6%	95.8%
Base / Ind1	352	83.2%	85.5%
Base / Ind2	73	88.2%	84.6%
Base / Ind3	110	83.6%	87.7%
Base / Ind4	39	84.6%	74.4%

Notes. [a]Percent accuracy for the base model was conducted using leave-one-out cross-validation and represents a baseline for how well the classifier was able to classify the same data on which it was trained. [b] Results based on k-nearest-neighbor classifier with k=8.

In the second set of prediction tests I combined the five data sets and used leave-one-out cross-validation to better estimate the prediction error that might be expected in generalizing the three decision styles to other IGT data sets. The results of the second set

of predictions tests are given in Table 4.5. Overall, the QDA model was able to predict the decision style of participants with 91.4%accuracy and the KNN model with 92.2% accuracy. I found no large and consistent differences in the prediction accuracies for each cluster. I believe these results serve as additional evidence supporting the external validity of our Chapter III results, and suggest that the prototype patterns associated with each decision style should generalize well as a basis for characterizing performance in the IGT.

Table 4.5 Leave-one-out tests of prediction accuracy.

Level	QDC mean accuracy	KNN mean accuracy
Overall	91.4%	92.2%
EV-LowFreq	87.9%	93.7%
EV-HighFreq	90.9%	91.4%
Frequency-Sensitive	95.1%	91.3%

Discussion

There are numerous reasons why a clustering solution found in one data set might not generalize to independent data sets. First, generalization might be challenged due to variability in the data itself across independent experimental collections, for example due to (i) variability in participant demographics and traits due to differences in recruiting and screening, (ii) variability in experimental procedures such as task instructions and the testing environment, and (iii) variability in incentives and motivation. Second, the inherent difficulties posed by unsupervised clustering pose technical challenges to the generalizability of solutions, for example due to: (i) overfitting a single sample of data, (ii) the sensitivity of clustering solutions to the choice of algorithm, dissimilarity measure and initialization, and (iii) the lack of a well-established, formal method for selecting the appropriate number of clusters. I addressed the latter set of challenges, and to some extent the former, using an ensemble clustering procedure combined with bootstrapping of the data set. To fully address the challenges posed by variability in data, I tested the external validity of the results from Chapter III by attempting to replicate them across a diverse set of four independent IGT data sets, and by assessing their ability to predict decision style across these data sets.

For each of five IGT data sets, I found a three-cluster solution, and for each of these solutions I found that the prototypical patterns of performance were remarkably similar. That the EV-Low, EV-High and Frequency-Sensitive patterns of performance were found across five data sets differing in mean age, gender ratio, collection location, and culture, suggests that these patterns represent fundamental differences in decision making behavior across individuals performing the IGT. The fact that the prevalence of each of these decision styles across experiments was found to be relatively stable further suggests that these differences are real rather than an artifact of either the clustering procedure or a particular data set. The results of the pairwise prediction tests showed that by representing each decision style as a group-averaged pattern of performance from a single data set (i.e. as a "cluster prototype"), I could reliably predict the decision style assigned to participants in other data sets. Lastly, the high level of generalization accuracy that was estimated using the combined data set of 885 participants provides support for the possibility of using "normed prototypes" of each decision style as a future way to analyze and perform inference in experimental and correlation studies of decision making using the IGT.

CHAPTER V. CHARACTERIZING INDIVIDUAL DIFFERENCES

Introduction

The clustering studies in the previous chapters revealed three robust groups of decision makers differing in their performance on the IGT. These groups were characterized in terms of choice performance as measured by differences in their learned preferences over the four decks. This characterization is necessarily incomplete, being derived only from an analysis of performance on the IGT itself. A more complete account of the differences across the groups must make contact with other concurrent and/or predictive measures, such as behavioral (for example, measures of performance on other cognitive tasks), latent (for example, psychometric measures of personality traits) or more fundamental measures (for example, electrophysiological measures or genetic correlates). Revealing associations between the three IGT decision groups and other measures not directly related to the IGT is a necessary aim to further characterize group differences. The purpose of the next set of studies was to investigate a more extensive set of measures that might reasonably be associated with differences in decision making in the IGT.

There is a small existing literature that has attempted to identify traits and cognitive measures associated with differences in decision making on the IGT (Table 5.1). The limited attention given to this area of research is somewhat surprising given the ubiquity of the claim that disadvantageous performance on the IGT reflects "risky decision making" and the stronger claim that this risky pattern of decision making is associated with impulsivity (2003). The set of measures investigated in this limited set of studies is far from complete, and those measures that have been studied have been found to be only weakly associated with performance on the IGT. Furthermore, multiple independent studies testing the same (or similar) measures have often produced conflicting results. Taken together, these studies suggest that women perform the IGT more disadvantageously than men and that performance on the IGT may be associated

with performance on the Wisconsin Card Sorting Task (WCST) as well as to measures of disinhibition, impulsivity, sensation, and reward seeking behavior (Table 5.1).

While it may simply be the case that associations between the IGT and previously studied traits and behavioral measures are either weak or non-existent, an alternative hypothesis is that true underlying associations may have been obscured by analyses that have implicitly assumed individual differences were well-captured by *%Good* as a dependent measure (either overall *%Good*, or the pattern of *%Good* by blocks) – an assumption shown to be incomplete by the results of the clustering analysis done in this dissertation. Specifically, it is possible that important concurrent measures might be associated with the three decision making groups identified in this dissertation, but produced weak or insignificant effects when tested against individual performance measured in terms of %Good (as in regression or correlation studies) or tested against groups defined in dichotomously in terms of advantageous/disadvantageous performance (as in analysis-of-variance studies). The first aim of the set of studies presented in this chapter was therefore to test for the degree of association between previously studied traits and cognitive measures and the three decision styles identified in this dissertation. By conducting analyses using the three groups as the focus of study, it is possible that stronger associations between traits and IGT performance might be revealed. A second and related aim was to investigate the relative merits of trait-based accounts of the IGT and more cognitive accounts. While trait-based accounts of the IGT have become widely accepted, another important hypothesis originally advanced by Farah and Fellows (Brand & Altstotter-Gleich, 2008; Brand, et al., 2007; Goudriaan, Grekin, & Sher, 2007; Sweitzer, et al., 2008; van Honk, Hermans, Putman, Montague, & Schutter, 2002) is that impaired decision making in the IGT is due to impairments in executive functions rather than to somatic biasing signals. By simultaneously investigating associations between both trait and cognitive measures and IGT group, I sought evidence that might favor one of these two competing accounts.

Table 5.1. Studies associating traits and cognitive measures with the IGT.

Measures	Identified association with IGT performance
Sex Differences	· No significant effects of sex on IGT performance (Bolla, et al., 2004; Denburg, et al., 2009; Overman, 2004; Reavis & Overman, 2001). · Significant effect of sex: men choose more advantageously than women (Reavis & Overman, 2001). · Men lower in testosterone more advantageous than those with higher testosterone (Goudriaan, et al., 2007). · Women have greater preference for low-frequency, positive expected value deck than men (Zermatten, et al., 2005).
Impulsivity (BIS11, UPPS, I7, DII)	· High scores on Premeditation scale of UPPS correlated with disadvantageous performance (Sweitzer, et al., 2008). No significant correlation on other three UPPS scales. · Higher BIS11 composite scores associated with more disadvantageous performance in the last block of the IGT (Franken, van Strien, Nijs, & Muris, 2008). · Higher scores on the I7 Impulsiveness scale associated with more disadvantageous performance(Goudriaan, et al., 2007). · No correlation between BIS11 total scale and IGT performance in substance dependent populations (Franken & Muris, 2005). · Higher scores on the Functional Impulsivity (FI) scale of the Dickman Impulsivity Inventory (DII) associated with more advantageous performance (Reavis & Overman, 2001).
Sensation Seeking (SSS)	· Higher SSS correlated with more advantageous performance. (Crone, Vendel, & van der Molen, 2003). · Higher scores on Disinhibition scale of SSS associated with more advantageous performance (Harmsen, Bischof, Brooks, Hohagen, & Rumpf, 2006). · No significant correlations between SSS scales and IGT performance in populations of cigarette smokers (van Honk, et al., 2002).
Behavioral Inhibition & Activation (BIS-BAS)	· Individuals with high BAS and low BIS performed more disadvantageously than individuals with low BAS and high BIS (Desmeules, et al., 2008). · Individuals with low BAS and high BIS performed more disadvantageously than individuals with low BIS and low BAS scores (Franken & Muris, 2005). · Higher scores on BAS Reward Responsiveness correlated with more advantageous performance. No significant correlations between BIS scales or other BAS scales and IGT (Suhr & Tsanadis, 2006). · Higher scores on BAS Fun Seeking scale and Reward Responsiveness associated with more disadvantageous performance (Brand & Altstotter-Gleich, 2008). · No significant association between the BIS Anxiety/Nervousness or Frustration/Tearfulness scales nor with the BAS Drive or Reward Responsiveness subscales (Reavis & Overman, 2001).
Depression (CES-D)	· No significant association between Center for Epidemiologic Studies Depression Scale (CES-D) and IGT performance (Denburg, et al., 2009). · No significant association between depression and IGT performance (Brand & Altstotter-Gleich, 2008).
Perfectionism (FMPS)	· The Doubts About Actions, Concern Over Mistakes and Personal Standards scales of the Frost Multidimensional Perfectionism Scale were not correlated with IGT performance (Lakey, Rose, Campbell, & Goodie, 2008).
Narcissism (NPI)	· Higher scores on the Narcissistic Personality Inventory were significantly associated with more disadvantageous performance on the IGT (Suhr & Tsanadis, 2006).
Affect (PANAS)	· Negative affect scale of PANAS associated with disadvantageous performance (Denburg, et al., 2009).
General Traits (EPQ-J, NEO-FFI)	· Higher scores on Neuroticism scale of the NEO Five Factor Inventory (NEO-FFI) was associated with more disadvantageous performance among older but not younger adults. No significant association was found between the Extraversion, Openness, Agreeableness, and Conscientiousness scales in either older or young adults (Hooper, Luciana, Wahlstrom, Conklin, & Yarger, 2008). · Higher scores on Neuroticism scale of Eysenck Personality Questionnaire-Junior (EPQ-J) were associated with more disadvantageous performance in adolescent males, but not females. Scores on the Psychoticism, Extraversion, and Externalizing scales were not significantly associated with IGT performance (Overman, 2004).
Executive Function (WCST, TOH, Dual-Tasking)	· Perseverative errors, non-perseverative errors, and trials-to-first-category on the Wisconsin Card Sorting Task (WCST) associated with more advantageous performance, particularly in the later stage of the task. · Performance on the WCST not associated with IGT performance in adolsescents (Brand, et al., 2007). · Performance on Tower of Hanoi (TOH) task not found to be associated with performance on the IGT (Hinson, Jameson, & Whitney, 2002; Jameson, Hinson, & Whitney, 2004). · Performance of a secondary serial order task interfered with performance on the IGT and anticipatory affective responses as compared to secondary tasks involving either verbal buffering or keyboard responses (Zuckerman, 1964, 1971; Zuckerman & Link, 1968).

Methods

To investigate possible associations between IGT decision style and demographic, trait and cognitive factors, the IGT was administered to 119 participants. In addition to the IGT, these same participants also performed two other cognitive tasks and completed a battery of personality assessments. Data from the IGT was clustered[11] using the ensemble clustering methods presented in Chapter III, and the association between the identified IGT decision style and the concurrent cognitive measures, self-reported trait measures, and demographic variables were then tested.

Procedures

Participants and study administration

All procedures were approved by the Institutional Review Board of the University of Michigan in Ann Arbor. The IGT data used in this study was collected from 119 participants. These participants were recruited from within the University of Michigan community as well as more broadly in the local Ann Arbor Community. Recruitment was done using flyers and postings on the Internet. All participants were screened for known psychiatric and neurological disorders. The screening questionnaire (see Appendix A) was made available to participants via the UM Lessons web-based questionnaire administration tool. The mean age of the participants was 21.73 (SD=6.16, range 17-54) and 56.36% were female. The study sessions lasted approximately 2 hours and consisted of 10 to 20 participants per session. Participants were paid based on their performance in the IGT component of the session. After participants provided informed consent, they participated in a session consisting of a 15 minute olfactory discrimination task (not reported here) followed by a self-paced set of personality questionnaires, the IGT, the Wisconsin Cart Sorting Task (WCST) and lastly a Digit Span task. Participants were free to take short breaks between each phase of the session, and the computer mandatory imposed breaks of two minutes in between each of the cognitive tasks. With the exception of the olfactory task, all questionnaires and tasks were administered by computer. Data from six of the 119 participants was excluded due to either a malfunction of the task administration software (one participant), failure to disclose prior experience

[11] This data set is the "Ind3" set that was used in the external validation study reported in Chapter IV.

with the IGT (four participants), or disclosure of unreported screening information after administration of the task (one participant). Data from the remaining 113 participants were then input into the ensemble clustering procedure described in Chapter III. Three participants were found to be outliers (more than three standard deviations from the group mean using z-scored Mahalanobis distance as a metric). Data from these participants were removed, and the data from the remaining 110 participants were used as the basis for the two studies reported in this chapter. The instructions shown to participants for the personality questionnaire, IGT, WCST and Digit Span are given in Appendices B, C and D.

Questionnaires and trait assessments

After completing the olfactory discrimination task lasting approximately fifteen minutes, participants then completed a self-paced set of questionnaires administered via the computer-based survey administration tool. The order of the questionnaires was identical for each participant. The total time spent completing the questionnaires was recorded for each participant. The mean completion time was 35.4 minutes and the minimum and maximum times were 22.9 minutes and 63.8 minutes, respectively. The first questionnaire (Appendix E) included demographic questions as well as five "lifestyle" questions pertaining to the frequency of cigarette smoking, alcohol consumption, drug use, and gambling. The collection of these measures was motivated by frequent use of the IGT in the study of pathological and non-pathological substance dependence and gambling. In addition to these four measures, participants were also asked to indicate their average hours of nightly sleep as well as their level of mathematical skill.

The demographic and lifestyle questionnaire was followed by fourteen standardized personality assessments designed to measure constructs either previously studied directly in relation to the IGT and/or that are well-known to be associated with decision behavior more generally (Table 5.2). The constructs covered by this set of assessments included: (i) risk taking and sensation seeking, (ii) impulsiveness, (iii) behavioral inhibition and activation, (iv) maximizing and regret, (v) perfectionism, compulsiveness, and indecisiveness, (vi) general affect, and (vii) general temperament and personality. Several of these constructs were assessed by multiple instruments. For example, impulsiveness was assessed directly by the UPPS Impulsive Behavior Scale, the Barratt

Impulsiveness Scale, and the Dickman Impulsivity Inventory – and less directly by several of the subscales of assessments designed to measure other related constructs (e.g., the Disinhibition subscale of the Sensation Seeking Scale and the Disinhibition subscale of the General Temperament Survey). The reason for using multiple assessments was diversity. Rather than making arbitrary decisions about which of several widely used assessments might "best" measure a particular construct, I chose to include multiple instruments that putatively measure different facets of the same construct. Two challenges in using multiple assessments of the same construct were the concomitant increase in dimensionality of the data to be analyzed, and the possibility of high correlations (and possibly multicollinearity) among sets of measures expected to be highly similar. These challenges were addressed by using dimensionality reduction methods, described in the Data Analysis section that follows. The simultaneous investigation of a wide range of measures comes with the associated pitfall of multiple comparisons, but I believed it more principled to acknowledge the exploratory nature of the study and to simultaneously consider a large set of relevant measures, rather than to partition this larger study into individual, independently reported studies for the purpose of avoiding the statistical onus of corrections for multiple comparisons. Furthermore, by using prediction models as a complement to hypothesis testing procedures, effects possibly obscured by statistical tests might be revealed.

Table 5.2 Summary of assessments administered to participants.

Assessments (abbrev.; items)	Scales collected	Source
Sensation Seeking Scale (SSS; 40)	Thrill and Adventure Seeking (TAS) Experience Seeking (ES) Disinhibition (DIS) Boredom Susceptibility (BS)	(Blais & Weber, 2006; Weber, Blais, & Betz, 2002)
Domain-Specific Risk Taking (DOSPERT; 50)	Financial (FIN) Health/Safety (HS) Recreational (REC) Ethical (ETH) Social (SOC)	(Dickman, 1990)
Dickman Impulsivity Inventory (DII; 23)	Functional Impulsivity (FI) Dysfunctional Impulsivity (DI)	(Whiteside & Lynam, 2001; Whiteside, Lynam, Miller, & Reynolds, 2005)
UPPS Impulsive Behavior Scale (UPPS; 45)	Urgency (URG) Premeditation (PRE) Perseverance (PER) Sensation Seeking (SEN)	(Barratt, 1985; Patton, Stanford, & Barratt, 1995)
Barratt Impulsiveness Scale (BIS-11; 30)	Attentional (ATT) Motor (MOT) Nonplanning (NPL)	(Carver & White, 1994)
Behavioral Inhibition and Activation System (BIS-BAS; 20)	Behavioral Inhibition (BIS) BAS – Drive (DRV) BAS – Fun Seeking (FS) BAS – Reward Responsiveness (RR)	(Schwartz, et al., 2002)
Regret and Maximizing Scale (RMS; 18)	Regret (R) Maximizing (M)	(Frost, Marten, Lahart, & Rosenblate, 1990)
Multidimensional Perfectionism Scale (MPS; 35)	Concern Over Mistakes (CM) Doubts About Actions (DA) Personal Standards (PS) Organization (ORG) Parental Criticism (PC) Parental Expectations (PE)	(Kagan & Squires, 1985)
Compulsiveness Inventory (CI; 11)	Indecision and Double-Checking (IDC) Detail and Perfectionism (DP) Order and Regularity (OR)	(Frost & Shows, 1993)
Frost Indecisiveness Scale (FIS; 15)	Indecisiveness (IS)	(Watson, Clark, & Tellegen, 1988)
Positive and Negative Affect Scale (PANAS; 20)	Positive Affect (PA) Negative Affect (NA)	(Watson & Clark, 1992)
General Temperament Survey (GTS; 90)	Positive Temperament (PT) Negative Temperament (NT) Disinhibition (DI)	(John, Donahue, & Kentle, 1991)
Big Five Inventory (BFI; 44)	Openness (O) Conscientiousness (C) Extraversion (E) Agreeableness (AA) Neuroticism (N)	(Beck, Erbaugh, Ward, Mock, & Mendelsohn, 1961; Beck, Steer, Ball, & Ranieri, 1996; Beck, Steer, & Brown, 1996)
Beck Depression Inventory (BDI-II; 21)	Total Depression Score (BDI)	(Appollonio, et al., 2002; Della Sala, MacPherson, Phillips, Sacco, & Spinnler, 2001; Shallice & Evans, 1978; Spreen & Strauss, 1998a)
Cognitive Estimation Test (CET; 10)	Total Absolute Deviation (TAD)	(Shallice & Evans, 1978; Spreen & Strauss, 1998b)

Cognitive Estimation Test (CET)

While the relative involvement of implicit and explicit decision processes in the IGT continues to be debated, there is little reason to doubt that the task involves processing numerical quantities to some degree. Setting aside whether this processing is done implicitly or explicitly, performance on the task requires estimation (possibly of payoff magnitudes, variances, and frequencies) and therefore it is possible that individual differences in the IGT might be associated with differences in cognitive estimation abilities. To test this hypothesis, participants completed the Cognitive Estimation Test (CET) as part of the assessment phase of the experimental session (Shallice & Evans, 1978). In the CET, participants respond to a series of ten questions that require estimation of quantities such as the height of the Empire State Building, the length of the average necktie, and the flight speed of a commercial jet. The CET was originally developed to measure impaired executive function in frontal patients (Spreen & Strauss, 1998b), but its use has been extended to healthy populations. The CET is thought to test estimation abilities and problem solving strategies in tasks in which exact computations are not possible. Performance on the task has been shown to be only weakly correlated with other commonly used tests such as the Wisconsin Card Sorting Task and the Tower of Hanoi (PEBL, Mueller, 2008). The test is scored by comparing each response to a table of normed values which indicate a degree of deviation ranging from -2 (underestimation) to +2 (overestimation), with zero indicating an accurate response within a small range of the true answer. The score for the test is the Total Absolute Deviation (TAD), the sum of the absolute values of the deviation scores computed for each question.

Iowa Gambling Task (IGT)

IGT data were collected using a computer-based administration of the IGT implemented in the Psychology Experiment Building Language (Bechara, 2007). This implementation of the IGT was identical in both graphics and sounds to the computer-based version of the IGT developed and used by Bechara and colleagues. I used the A'B'C'D' version of the task, the same version used in the commercially available IGT assessment tool (Grant & Berg, 1948; Spreen & Strauss, 1998b), and in the collection of each of the data sets presented in Chapter IV, with one exception: forty additional cards

were added to the decks to avoid deck depletion as was found to have occurred in prior administrations of this task. These additional cards were generated randomly based on the same rules for determining payoff magnitudes and frequencies used to generate the cards in the 60-card version of this task. For convenience, in the remainder of this chapter I will omit the apostrophes and refer to the four decks simply as *A*, *B*, *C*, and *D*. Participants were paid based on their performance on this task. The payoff schedule was provided to participants in both the consent form and in the instruction sheet given out prior to performing the task. The minimum payment was $5.00 and corresponded to IGT net profits of -$4000 or less. The maximum payment was $50.00 and was awarded for IGT net profits of $4000 or more. There were eight additional levels of payoffs associated with IGT net profits ranging from -$4000 to +$4000 in increments of $1000. The primary dependent measure used in this study was group membership as computed using the ensemble clustering procedure described in Chapter III. The percentage of cards selected from the two advantageous decks (*%Good*) was also computed to confirm proper administration of the task. The decision time on each trial was also collected.

Wisconsin Card Sorting Task (WCST)

Given the currently inconclusive evidence relating IGT and WCST performance, I was interested in further testing this relationship as mediated through analysis at the level of the group differences in the IGT. After completing the IGT, participants performed a computer-based version of the WCST implemented in PEBL. The WCST is known to require a range of executive functions including set shifting, rule abstraction, response inhibition, and feedback utilization (Greve, Stickle, Love, Bianchini, & Stanford, 2003). In the WCST, participants are given a set of cards which they must place, one at a time, on one of four decks according to a sorting rule which must be learned through trial-and-error. Each card depicts one or more identical geometric objects and the cards differ in the shape (circle, triangle, square, cross), color (red, green, blue, yellow), and number (1, 2, 3, 4) of objects shown. At any given time, a single sorting rule is in force (number, color, shape) and critically, after successfully sorting some number of cards, the rule changes. Participants must recognize via feedback that the rule has changed, and must learn each new rule through trial-and-error. The WCST used in this study was the original 128 card version with the sorting rule changed after a run of ten correct responses. The task was stopped after participants sorted 128 cards, or successfully

completed nine categories. The dependent measures obtained from this task were: Categories Completed (CC), Number of Trials (NT), Perseverative Errors (PE), Non-Perseverative Errors (NPE), Trials to First Category (TFC), the mean length of perseverative runs (MeanPR), and the length of the longest perseverative run (maxPR). While each of these dependent measures is thought to capture somewhat different aspects of WCST performance, it has been argued that the performance by healthy adults is best captured by a single factor (Miller, 1956). The seven measures were thus collapsed into a single WCST score for each participant using Factor Analysis as will be described in the following section. In addition to the accuracy and error measures, the decision time on each trial was collected.

Digit Span Task

The Digit Span (Lilliefor, 1967) is known to test working memory capacity as well associated abilities in verbal rehearsal and was included as a cognitive task to test the possibility that differences in IGT performance might be related to differences in working memory capacity and/or executive rehearsal processes. In the Digit Span, participants are presented with lists of numbers, one number at a time, and are instructed to remember each digit in a list, in the order in which they are presented. In the forward version of the task, the length of the list is increased until a stopping criterion is met. After completing the WCST, participants performed a computer-based version of the forward Digit Span task as the final step of the experimental session. The Digit Span was implemented in PEBL. Digits were presented visually on the computer monitor and also simultaneously spoken aloud via a recorded voice heard through closed-type headphones. Because the sessions consisted of multiple participants, it was not feasible to collect auditory responses. Instead, participants responded by pressing the number keys located at the top of their computer keyboard. The list size was increased from three to a maximum of fifteen digits, with three trials presented at each list length. The task was stopped when a participant failed to successfully complete two of three trials at the current list length. The dependent measure was the digit span, the largest list length at which a participant completed two out of three trials. Response time on each trial was also collected.

Data Analysis

Data processing

The dependent measures for the IGT, WCST, and Digit Span tasks were computed directly from the raw data using MATLAB® (2008b, The MathWorks, Inc.: Natick, MA). Unless noted otherwise, all analyses reported in the remainder of this chapter were conducting using MATLAB. The data collected from the fifteen assessments consisted of 420 responses for each of the 110 participants. Scores on each of the 45 trait scales and the TAD score for the CET (c.f., Table 5.2) were computed from these responses using an automated assessment toolbox developed in MATLAB by the author. The five scores on the lifestyle measures (Smoking, Drinking, Drug Use, Gambling, and Sleep) and the self-reported Math ability score were computed directly from individual responses. To determine the reliability of the trait data, Cronbach's alpha statistic of internal-consistency reliability was computed for each scale and compared to published norms as well as to the widely accepted heuristic that values in the range 0.60 - 0.70 are indicative of acceptable reliability and values of 0.80 or higher are indicative of good reliability.

Each of the cognitive and trait measures was checked for normality using a two-sided Lilliefors test (Hastie, et al., 2001; Rencher, 2002). Measures found to significantly depart from normality were transformed using the following automated procedure. First, the skew of each measure was computed and those with a negative skew were reflected. Transformations were then applied stepwise in order of increasing strength (square-root, then logarithmic, then inverse), with the Lilliefors test repeated on each step. If a transformed measure was accepted as normal according to the normality test, the procedure was stopped. If no transform led to normality, the transform producing the smallest absolute skew was used. After transformation, all measures that were reflected were once again reflected to preserve the original interpretation of their values. Lastly, because the units of measure were not commensurate across the set of measures, all measures were standardized.

Dimensionality reduction of trait data

Given the prior expectation that many of the trait scores would be highly correlated due to the use of multiple instruments for each of the major constructs that were assessed, a factor analysis (using principal components extraction) of the data correlation matrix was conducted to identify the most important latent structure in the data and to

reduce the dimensionality of the data prior to further analysis (Rencher, 2002). In typical applications of exploratory factor analysis to personality assessment, the primary objective is to reveal previously unidentified latent factors based on item-level responses. Exploratory factor analysis is often contrasted with confirmatory factor analysis which assumes prior knowledge of the latent factors in the data and seeks to test hypotheses related to these factors. The trait measures analyzed in this study were factor scores previously identified and validated in the development of the assessments used to generate them. As such, the application of factor-analytic methods fell somewhere between the exploratory and confirmatory methods and several considerations needed to be taken into account to appropriately factor-analyze these data.

First, as previously mentioned, the most similar among the trait measures were expected to be highly correlated, for example the Negative Temperament scale of the GTS and the Negative Affect scale of the PANAS. These highly correlated measures were expected to load on a common factor. However, other related measures were originally developed specifically to identify unique facets of the constructs they purport to measure, as for example in the case of the Functional and Dysfunctional measures of impulsivity in the DII as compared to the Attentional, Motor and Nonplanning measures of the BIS11. Thus, while the correlations between these related measures were expected to be greater than their correlations with unrelated measures, it was also expected that the presence of these measures would produce a very flat eigenvalue decomposition of the data which would in turn lead to overestimation of the number of factors using the standard approach of selecting the number of factors based on the number of eigenvalues greater than one. The practical significance of this expectation was that extra care needed to be taken in selecting the appropriate number of factors to represent the trait data. Two procedures were used to identify the number of factors to use in representing the trait data. The scree plot of the eigenvalues extracted using principal components analysis was examined and the number of supported factors was determined based on the inflection point ("knee") in the plot. A second procedure was accomplished by (i) examining the loadings of each factor as additional factors were added to the solution being fit, and then (ii) selecting the number of factors at the point when the addition of another factor resulted in a "low quality" factor being generated. In factor analysis, factor quality is often indicated by the presence of several highly

loaded variables: factors with only small loadings or with only a single highly loaded variable are typically not interpreted. Using this concept of quality, I varied the number of factors from 2 to 20 and for each factor solution, identified the number of high-quality factors for each solution using the criteria that a high-quality factor should have at least three variables loaded with a correlation of at least 0.50. The smallest solution for which every factor was determined to be high-quality was then selected. A final consideration in appropriately applying factor analysis to the data was the stringency with which factors loadings were interpreted. In exploratory factor analysis, interpretation of factors is typically on loadings that exceed than 0.30; given the substantial *a priori* knowledge of the latent structure expected from the data in this study, the criterion for interpretation in this study was set at 0.50. Prior to interpreting the selected factor solution, the extracted principal components were rotated using orthogonal Varimax rotation (Rencher, 2002). After selecting and interpreting a factor solution, scores on each factor were generated for each participant.

Dimensionality reduction of WCST data

The seven error and accuracy measures obtained from the WCST were collapsed into a single WCST score for each participant using factor analysis with principal-components extraction and orthogonal Varimax rotation. This single WCST score was then used in all subsequent analysis that included the WCST. The response time measure was not included in this factor, but analyzed independently along with response time measures from the other cognitive tasks.

Tests of association and prediction

To investigate the primary aims of this study, traits and concurrent cognitive measures were tested for mean differences across each of the three groups of IGT decision makers. The measures were divided into four sets and these sets were analyzed separately to determine their relative importance as possible correlates of IGT decision group. These sets were: lifestyle measures, trait measures (as represented by the results of the factor analysis), cognitive measures, and response time measures. To investigate the degree of association between each set of measures and IGT decision group, I conducted multivariate analyses of variance (MANOVAs) with IGT decision group as the between-subjects factor and sex as a covariate. Statistical significance was determined

and will be reported based on Wilks' Λ and Pillai's Trace, statistics known to be robust in the presence of unequally sized groups, as is the case with the three IGT decision groups to be tested (Tabachnick & Fidell, 2001). Positive multivariate omnibus tests for mean differences were followed by univariate analyses of variance (ANOVAs) for each measure based on Bonferoni adjusted p-values to account for multiple comparisons within a given set of measures. Partial η^2 values were used to evaluate the degree of association between IGT groups and measures found to significantly differ across the groups.

Given the number of measures tested in this study, it was possible that important associations might be present but fail to reach significance due to the use of corrected p-values and/or lack of power due insufficient sample size, high variability, or both of these factors. As a complement to hypothesis testing procedures, a classifier model was used to further investigate the possible associations between the traits and cognitive measures and the IGT decision groups. The four sets of measures used in the MANOVA analyses were each entered as sets of predictors in a series of multinomial logistic regression models which were used to predict the group membership of each participant. Like multiple regression models, logistic regression models fit a set of beta values representing the coefficients of each predictor in a linear equation used to predict an outcome variable. In logistic regression, the outcome is nominal rather than continuous and may take any number of discrete values. The regression equations model the log-odds of a participant being in one group rather than in a reference group. The choice of reference group is arbitrary, and a final model consists of one logistic regression equation for each of the groups not used as the reference group (Spreen & Strauss, 1998b).

To evaluate the association between measures of interest and the IGT decision groups, logistic regression models were fit to each of the four sets of measures (lifestyle, trait, cognitive, and response times) and the significance of each model was determined using a χ^2-test of the log-likelihood of each model compared to the log-likelihood of a reduced model with no predictors and only an intercept term. The relative importance of models was then evaluated by determining the overall accuracy with which each set of measures could predict group membership for each of the participants. Finally, the relative contribution of predictors within each model was evaluated directly based on exponentiated beta values which indicate the increase in the log-odds of membership in a group (relative to the reference group) for each unit change in a predictor;

exponentiated beta values were examined for all predictors found to be significant based on the Wald test.

Results

Data Validation

After processing the trait assessments, the internal-consistency reliabilities were checked. Of the 45 trait measures collected, 22 (49%) had reliabilities greater than 0.80, 12 (27%) had reliabilities between 0.70 and 0.80, 9 (20%) had reliabilities between 0.60 and 0.70, and 2 (4%) had reliabilities less than 0.60 (Appendix F gives the reliabilities and published norms for each measure). The reliability of the Attentional scale of the BIS11 was 0.59, but this measure was retained because this value was higher than the published norm of 0.58 for this scale. The reliability of the Boredom Susceptibility scale of the SSS was 0.54 and although slightly lower than the published norm of 0.57, this difference was not deemed sufficient to merit eliminating the scale from further analysis. The reliability of the Cognitive Estimation Test included within the trait assessments was found to be 0.48 which is substantially higher than a college-age norm of 0.37 for this task (Rencher, 2002).

Results from the WCST and Digit Span tasks were checked to ensure that administration of these tasks was not anomalous. Results for the WCST and Digit Span tasks are shown in Table 5.3. Overall, participants performed well on the WCST, completing a median number of categories of 8 out of the 9 possible and sorting on average 80% of the cards correctly. Three participants failed to understand the WCST as evidenced by their failure to complete a single category correctly; these participants performed within the normal range on the IGT and Digit Span. The WCST data from these three participants were excluded from further analysis. No anomalous results were found for the Digit Span task; participant spans ranged from 4 to 10 with a mean span of 7.11.

Table 5.3. Summary of performance on the WCST and Digit Span.

Measure	WCST CC [a]	WCST NT	WCST PE	WCST NPE	WCST TFC	WCST Mean PR	WCST Max PR	Digit Span
Mean	7.57	123.83	15.58	9.21	13.51	1.99	4.08	7.11
SD	1.78	6.63	6.32	6.78	5.51	0.73	2.11	1.38

Notes. Categories Completed (CC); Number of Trials (NT); Perseverative Errors (PE), Non-Perseverative Errors (NPE), Trials-to-First-Category (TFC), Mean and Maximum Length of Perseverative Runs (MeanPR, MaxPR).

To confirm that the administration of the IGT was consistent with expected results, the *%Good* and *%EEV* measures was computed for each participant and compared to the results from another large, independent data set (the Base data set presented in chapter IV). There were no significant differences across the two data sets for either measure: $t(426)=0.5938$, $p>0.10$ for *%Good* and $t(426)=-0.4390$, $p>0.10$ for *%EEV*.

The results of applying the ensemble clustering procedure to the data were previously presented in chapter IV. The data supported a three cluster solution with 40%, 35%, and 25% of the participants assigned to the Frequency-Sensitive, EV-LowFreq, and EV-HighFreq groups, respectively. The distribution of participants across the groups was compared to the Base data set and no significant difference was found: $\chi^2(2)=1.08$, $p>0.10$.

Before using IGT group for the primary analyses of this study, I tested for differences in sex, age, handedness, and education across the groups using independent-sample *t*-tests or χ^2-tests, as appropriate. There were no significant differences in any of these variables across the groups: sex $\chi^2(2)=1.26$, $p>0.10$; age $F(2,106)=2.42$, $p>0.10$; handedness $\chi^2(4)=6.08$, $p>0.10$; education $F(2,107)=0.29$, $p>0.10$. Age, handedness and education were therefore excluded from all subsequent analyses. Although sex was not found to be significantly different across the groups, there are known sex-differences in many of the traits assessed in this study, and therefore sex was included as a covariate in all subsequent analyses.

Factor Analysis of Trait Data

Kaiser's Measure of Sampling Adequacy (MSA) provides an indication of whether or not a correlation matrix might have structure sufficient to produce satisfactory results from factor-analysis, with a MSA values larger than 0.8 considered to be positive support (Bechara, 2004). The MSA for the trait data considered in this study was 0.83. Principal-

components extraction was performed on the set of 45 trait measures and the resulting eigenvalues are shown in the scree plot depicted in Figure 5.1. The scree plot is relatively flat across the majority of the components, as was expected given the fact that the data consisted of a set of measures that were previously derived via factor-analytic methods. The large incremental change in the eigenvalues from the first to the fourth and the subsequent smaller incremental changes following the fourth eigenvalue suggest a four component solution as a parsimonious representation of the data in reduced dimensions. Taken together, the first four components account for 49% of the variance in the data. It is worth noting that for these data, the standard procedure for determining the number of factors based on the number of eigenvalues greater than 1 would have selected a twelve-factor solution.

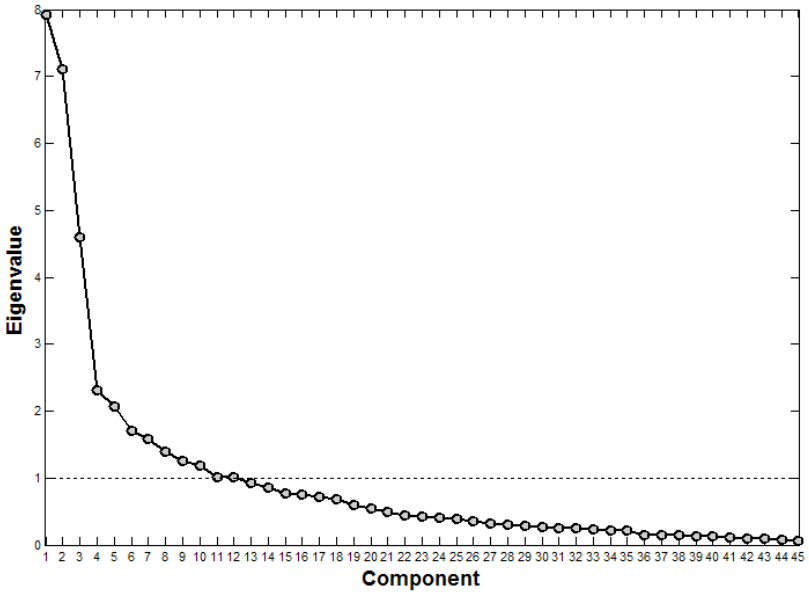

Figure 5.1. Scree plot of eigenvalues from principal-components analysis performed on correlation matrix of the 45 trait measures. As expected, the plot is relatively flat as a result of analyzing a set of measures that were themselves previously derived via factor-analysis.

As a second approach to selecting the number of factors, the factor-quality procedure outlined in the Methods section was performed on the trait data. The results of this analysis are shown in Figure 5.2. Conceptually, if each factor added to an initial two-factor solution produced a high-quality factor, the plot should follow the diagonal line shown in the figure; the point at which the plot departs and fails to return to the diagonal would indicate that larger models do not necessarily produce better solutions.

Applied to the trait data, this selection procedure suggested that models fit with more than four or five factors did not produce parsimonious solutions. The fact that five factor model itself produced only four factors meeting the quality criteria favors the four-factor solution as more parsimonious. Taken together, the scree test and the quality test both provided support for a four-factor solution.

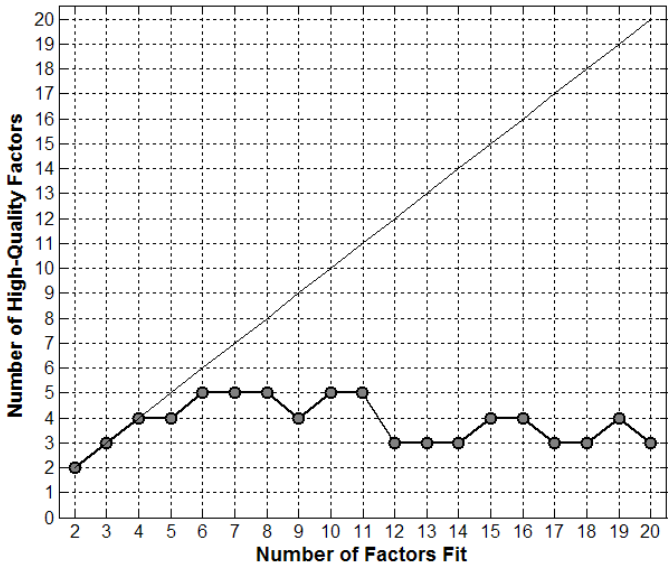

Figure 5.2. Results of factor quality procedure applied to the trait data. As factors are added to an initial two-factor model (horizontal axis), the number of factors meeting the quality criteria were identified (vertical axis). An idealized plot should follow the diagonal, and then depart at a point when larger models are no longer parsimonious. For the trait data shown in the figure, the plot departs from the diagonal after four or five factors, suggesting that larger models were not parsimonious. The five factor model generated only four factors meeting the quality criteria, and therefore the four-factor model was selected.

A factor analysis using principal-components extraction and Varimax rotation was then performed on the data, with the number of factors to be modeled fixed at four. The loadings (greater than 0.50) and communalities (h^2) for the four-factor solution are given in Table 5.4. The full rotated loading matrix is provided in Appendix G. The factor analysis produced a clear, simple result structure, as evidenced by: (i) high correlations (loadings) between measures and factors, (ii) the large number of measures highly loaded on each factor, (iii) the fact that only one variable highly loaded on multiple factors (Neuroticism), (iv) the loading of related measures in the same factor, and (v) the generally high proportion of common variance component of each measure accounted for by the factors (h^2). Examination of the loadings matrix shown in Table 5.4 led to the

following interpretation of the four factors. The first factor ("Positive-Drive") was associated with assertiveness and enthusiasm, positive affect and outlook, highly goal-driven behavior, and functional impulsivity contrasted with low levels of indecision. The second factor ("Dysfunctional Impulsivity") was highly correlated with measures from each of the three impulsivity assessments (DII, BIS11, UPPS) as well as with the related constructs of disinhibition and lack of conscientiousness. The third factor ("Negative Inhibited Perfectionism") was associated with perfectionism and compulsivity, maximizing and regret-oriented decision making combined with the general traits of negative temperament and neuroticism. The fourth factor ("Sensation Seeking") was correlated with self-reported risk-taking and sensation-seeking behavior as measured by the two instruments that assess these traits (the SSS and DOSPERT) as well by related scales from the UPPS and BIS-BAS. Each of the factors reflected known associations between the measures that loaded on them. The factor score coefficients produced by the factor analysis were used to generate scores for each participant, on each of the four factors. Scores on the 45 trait measures were thus reduced to scores on four trait factors and these were used to test the association between traits and the IGT decision groups. Higher scores on factors corresponded to greater association with each of the four traits represented by the factors.

Table 5.4 Rotated loading matrix from factor analysis of trait data.

Measure (Assessment)	Factor 1	Factor 2	Factor 3	Factor 4	h^2
Extraversion (BFI)	+ 0.74				0.56
Functional Impulsivity (DII)	+ 0.72				0.57
Positive Temperament (GTS)	+ 0.68				0.52
Positive Affect (PANAS)	+ 0.68				0.50
Drive (BIS-BAS)	+ 0.53				0.59
Indecisiveness (FIS)	– 0.63				0.53
Indecision/Double Checking (CI)	– 0.56				0.38
Dysfunctional Impulsivity (DII)		+ 0.71			0.71
Nonplanning Impulsivity (BIS11)		+ 0.70			0.58
Attentional Impulsivity (BIS11)		+ 0.67			0.54
Disinhibition (GTS)		+ 0.66			0.55
Motor Impulsivity (BIS11)		+ 0.63			0.63
Premeditation (UPPS)		– 0.54			0.49
Perseverance (UPPS)		– 0.62			0.61
Conscientiousness (BFI)		– 0.78			0.71
Concern Over Mistakes (FMPS)			+0.75		0.57
Doubts About Actions (FMPS)			+0.68		0.58
Regret (RMS)			+0.66		0.54
Negative Temperament (GTS)			+0.65		0.67
Maximizing (RMS)			+0.59		0.43
Personal Standards (FMPS)			+0.59		0.54
Behavioral Inhibition (BIS-BAS)			+0.57		0.55
Parental Expectations (FMPS)			+0.56		0.34
Neuroticism (BFI)	– 0.52		+0.55		0.61
Order/Regularity (CI)			+0.52		0.31
Organization (FMPS)			+0.52		0.47
Sensation Seeking (UPPS)				+0.81	0.72
Recreational Risk Taking (DOSPERT)				+0.80	0.72
Thrill & Adventure Seeking (SSS)				+0.74	0.58
Fun Seeking (BIS-BAS)				+0.68	0.62
Health & Safety Risk Taking (DOSPERT)				+0.58	0.48
Ethical Risk Taking (DOSPERT)				+0.56	0.49
Percent of Variance Explained	0.12	0.12	0.14	0.11	-
Cumulative Percent of Variance Explained	0.12	0.24	0.38	0.49	

Notes. Measures shown in italics are negatively correlated with their associated factor.

Factor Analysis of WCST Data

The results from the factor analysis performed on the WCST data are shown in Table 5.5. As intended, the analysis produced a single aggregate WCST factor based on a weighted combination of the seven original WCST measures. The number of categories completed, the number of trials completed and the two types of errors contributed somewhat more to the aggregate factor than two measures of perseverative runs. The aggregate WCST factor accounted for 52% of the variance in the raw data. The factor score coefficients produced by the factor analysis were used to generate a single WCST score for each participant, with higher scores representing better performance on the task.

Table 5.5 Rotated loading matrix from factor analysis of WCST data.

Measure	Factor 1	h^2	Score Coefficients
Categories Completed (CC)	+ 0.89	0.79	+0.245
Number of Trials (NT)	– 0.83	0.68	–0.228
Perseverative Errors (PE)	– 0.83	0.69	–0.229
Non-Perseverative Errors (NPE)	– 0.75	0.56	–0.207
Trials to First Category (TFC)	– 0.67	0.20	–0.125
Maximum Perseverative Run (MaxPR)	– 0.51	0.26	–0.139
Mean of Perseverative Runs (MeanPR)	– 0.45	0.44	–0.184
Percent of Variance Explained	0.52	-	-

Association between Measures and IGT groups

Mean differences in the lifestyle, trait, cognitive and response time measures across IGT groups were tested using full-factorial MANOVAs with IGT Group and Sex entered as independent variables in all tests. All univariate tests and post-hoc comparisons following the MANOVAs were performed with Bonferroni corrected significance levels, unless otherwise noted.

Lifestyle measures

The multivariate omnibus test using the five lifestyle measures (Smoke, Drink, Drugs, Gamble, Sleep) as dependent variables found no significant difference in the set of measures across IGT Group, (Wilks' $\Lambda(10,200)=0.971$, $p>0.9$; Pillai's Trace$(10,202)=0.30$,0 $p>0.9$). A significant sex-difference was found (Wilks' $\Lambda(10,200)=0.854$, $p<0.01$; Pillai's Trace$(10,202)=0.146$, $p<0.01$), but there was no significant interaction between IGT Group and Sex (Wilks' $\Lambda(10,200)=0.933$, $p>0.7$; Pillai's Trace$(10,202)=0.068$, $p>0.7$).

Univariate tests revealed no significant differences in any of the five lifestyle measures across the IGT groups; uncorrected comparisons were also not significant.

Trait measures

Next, a MANOVA was performed using the four trait factors as the dependent measures. No significant association we found between the four trait factors and either IGT cluster (Wilks' $\Lambda(8,202)=0.931$, $p>0.5$; Pillai's Trace$(8,204)=0.070,0$ $p>0.5$) or Sex (Wilks' $\Lambda(4,101)=0.946$, $p>0.2$; Pillai's Trace$(4,101)=0.054$ $p>0.2$). The interaction between IGT group and Sex was also not significant. Univariate tests revealed no significant differences in any of the four trait factors across IGT Group. No univariate post-hoc comparisons of the four traits across IGT groups were significant (or approaching significance); uncorrected comparisons were also not found to be significant.

Cognitive measures

I next investigated the association between the four cognitive measures (WCST, Digit Span, CET, and Math) and IGT Group. This test revealed a significant association between the set of cognitive measures and IGT group (Wilks' $\Lambda(8,196)=0.837$, $p<0.05$; Pillai's Trace$(8,198)=0.168$, $p<0.05$) as well as Sex (Wilks' $\Lambda(4,98)=0.890$, $p<0.05$; Pillai's Trace$(4,98)=0.110,0$ $p<0.05$). The interaction between IGT Group and Sex was not significant in the omnibus test. Univariate tests using each of the four measures revealed that only the aggregate WCST score was significantly associated with IGT Group and furthermore that this measure also differed significantly across Sex (Table 5.6). The interaction between IGT Group and Sex for the aggregate WCST score was not significant. Consistent with the literature on sex-differences in the WCST, women performed better than men with a mean difference of 0.442 standardized units on the aggregate WCST measure, but this was a small effect (partial $\eta^2=0.05$). Of primary interest in this study were the differences in WCST performance across the three IGT groups. Post-hoc tests revealed that the Frequency-Sensitive participants exhibited significantly poorer performance on the WCST than participants in the EV-LowFreq group (mean difference of 0.641 standard units, $p<0.05$) as well as the EV-HighFreq group (mean difference of 0.61 standard units, $p<0.05$). There was no significant difference in WCST performance between the EV-LowFreq and EV-HighFreq groups.

Table 5.6 Univariate tests of cognitive measures versus IGT Group and Sex.

Measure	IGT Group			Sex		
	F	p	Partial η^2	F	p	Partial η^2
WCST Aggregate Score	5.801	0.004	0.103	5.444	0.022	0.051
Digit Span	1.592	0.209	0.031	0.021	0.886	0.000
Cognitive Estimation Test (CET)	.995	0.373	0.019	0.288	0.593	0.003
Math Ability (self-reported)	.887	0.415	0.017	6.299	0.014	0.059

To better understand which measures of WCST performance contributed to the poorer performance found for the Frequency-Sensitive group, pair-wise mean differences across the IGT Groups were tested using independent sample *t*-tests (Table 5.7). Sex was not considered because the IGT Group x Sex interaction for the WCST was already found to be non-significant. To facilitate interpreting the mean differences, raw WCST scores were used rather than the standardized and transformed measures as were used in the previous analyses. Participants in the Frequency Sensitive group performed significantly worse on the IGT compared to participants in the two EV-sensitive groups. In particular, the Frequency-Sensitive participants on average completed about one less category, terminated the task 3-4 trials later, and made 3-4 more perseverative errors with longer runs of perseverative choices. Interestingly, participants in the Frequency-Sensitive group did not differ from the other groups in non-perseverative errors or the number of trials required in learning the first category.

Table 5.7 Mean differences in WCST measures by IGT Group.

Measure	Frequency - Sensitive vs. EV-LowFreq	Frequency-Sensitive vs. EV-HighFreq	EV-HighFreq vs. EV-LowFreq
Categories Completed (CC)	-0.90**	-0.80	-0.10
Number of Trials (NT)	3.70***	3.11**	0.58
Perseverative Errors (PE)	4.25***	3.24*	1.00
Non-Perseverative Errors (NPE)	1.63	2.09	-0.46
Trials to First Category (TFC)	0.99	1.95	-0.96
Maximum Perseverative Run (MaxPR)	1.19**	1.67***	-0.48
Mean of Perseverative Runs (MeanPR)	0.40**	0.45**	-0.05

Note. $^{*}=p<0.10$, $^{**}=p<0.05$, $^{***}=p<0.01$

Response time measures

I next investigated the association between response time measures and IGT Group. The five response time measures entered as dependent variables in the MANOVA were (i) mean response time in the IGT, (ii) mean response time in the WCST, (iii) mean response time in the Digit Span, and (iv) the total time participants spent filling out the trait assessments. This test revealed no significant association between these measures and IGT group (Wilks' $\Lambda(8,198)=0.947$, $p>0.7$; Pillai's Trace$(8,200)=0.054$, $p>0.7$) nor with Sex (Wilks' $\Lambda(4,99)=0.983$, $p>0.7$; Pillai's Trace$(4,99)=0.017$,0 $p>0.7$). The interaction between IGT Group and Sex was not significant in the omnibus test. Univariate tests revealed no significant differences in any of the response time measures across the IGT groups; uncorrected comparisons were also not significant.

Univariate ANOVA on Individual Measures

While the test of association between the trait factors and IGT group failed to show significant associations, it was possible that in reducing the 45 individual trait scales to four factors, important trait correlates may have been obscured. Although the failure to find significant associations was not surprising given that strong correlations have not been found with any consistency in the existing literature on the IGT, I nevertheless wanted to further confirm the results of the MANOVA tests. I therefore conducted exploratory, post-hoc analyses directly on each trait measure using univariate ANOVAs with IGT group as the independent variable. I also examined the patterns of trait scores across the three groups to determine whether or not there were regularities in their distributions (Figure 5.3). The plots in the figure show the standardized scores (Z values) for each trait, across each of the three IGT decision groups. The traits are organized in related sets. The related traits of compulsivity, perfectionism, maximizing and regret, and indecision are located in the bottom third of the figure; the impulsivity traits and the risk-taking and sensation-seeking traits in the middle of the plots, and the measures of general personality and affect are located in the top third of the plots. While there seem to be some regularities across the IGT groups in scores on related measures (for example, in the pattern of scores for the related set of compulsivity, perfectionism, regret/maximizing, and indecision traits), what is most evident from the figure is that the differences in trait scores across the three groups were very small, with even the largest difference (the DI scale of the DII) differing by only half of one standardized score. The

data shown in the figure offer little support for the hypothesis that the failure to find significant differences was due to the aggregation of individual trait scores into the four trait factors.

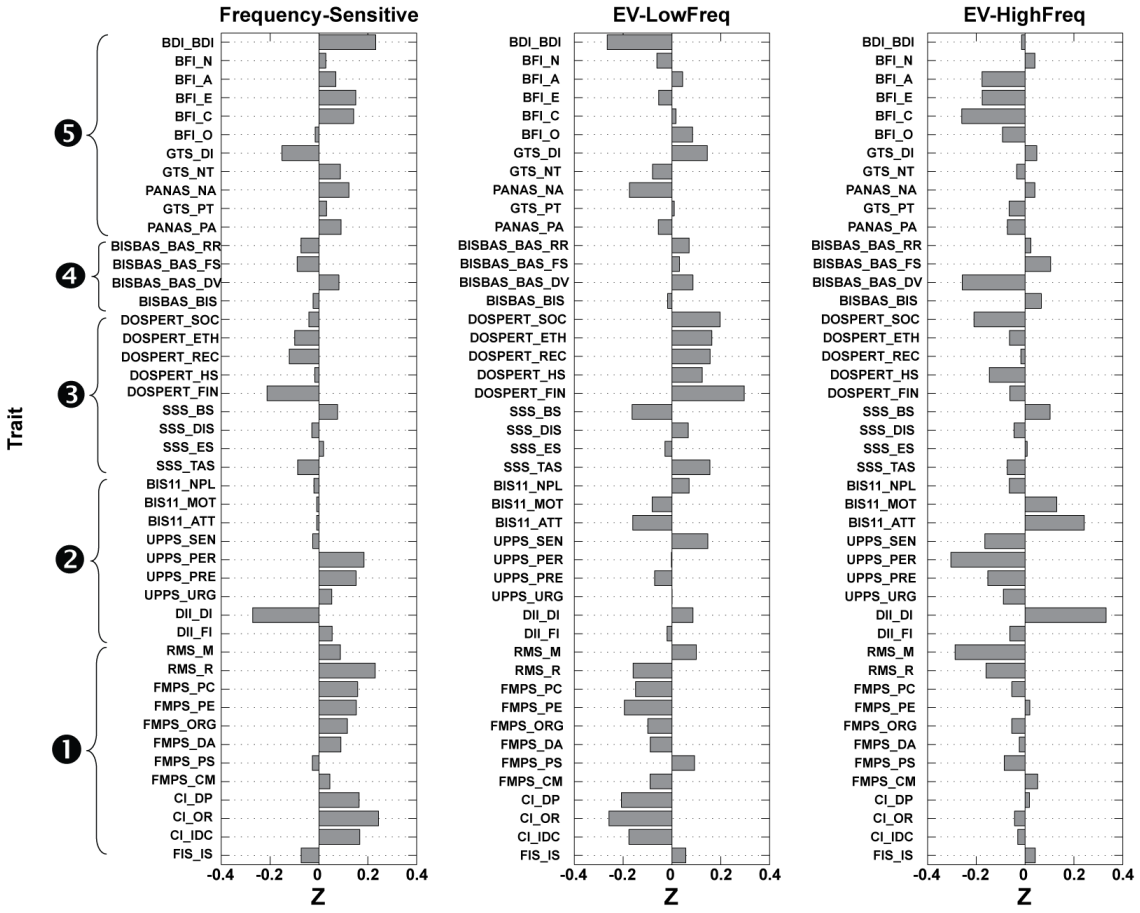

Figure 5.3 Mean standardized trait scores for each of the three IGT decision groups. While there seem to be differences across the groups among sets of related traits, these differences are very small in terms of standardized scores. These data do not suggest that significant differences in individual trait scores were obscured through the use composite scores obtained through factor analysis. Key to sets of measures: (1) Compulsiveness and Perfectionism, (2) Impulsiveness, (3) Sensation Seeking and Risk Taking, (4) Behavioral Inhibition and Activation, (5) General personality and affective traits. See Table 5.2 for trait abbreviations.

Figure 5.4 shows the results of one additional analysis done to investigate the possibility that important trait associations might have been obscured by the use of the four-factor representation of the trait data. The figure gives the uncorrected p-values generated by univariate tests of mean differences across the three IGT groups for each of the 45 traits. Confirming the results suggested by the analysis shown in Figure 5.3, none

of the traits was found to be significant except the Dysfunctional Impulsivity scale of the DII; few traits approached even marginal significance. Furthermore, among the set of six traits with the smallest p-values, there was no obvious pattern: two were impulsivity traits (DII_DI, UPPS_PER), one a risk taking trait (DOSPERT_FIN), one a compulsiveness trait (CI_OR), one the regret trait (RMS_R), and one a measure of depression (BDI_BDI). There was therefore little evidence of a robust pattern of associations provided by this additional analysis.

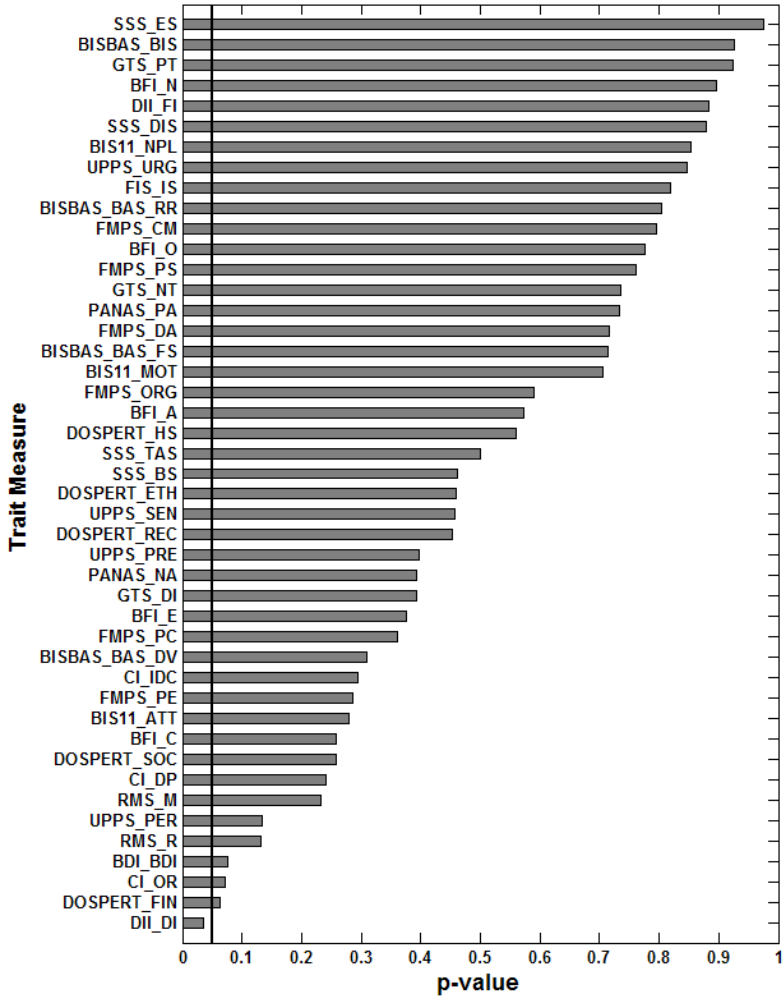

Figure 5.4 Results from univariate testsfor mean differences in each of the 45 trait measures across the three IGT decision groups. The level of significance (uncorrected p-values) is shown for each measure. Only one measure (the Dysfunctional Impulsivity scale of the DII) is significant at an uncorrected level of 0.05, and few traits even approach marginal significance. See Table 5.2 for trait abbreviations.

Prediction Results

The multivariate tests failed to reveal significant associations between trait factors and IGT decision group, and the exploratory post-hoc tests using individual trait measures revealed only one measure that was significant at an uncorrected alpha level of 0.05. One possible limitation in these tests was the size of the sample (N=110). While the results presented in the preceding sections offer little support for the presence of strong associations between any trait measures and IGT decision group, it was possible that one or more of these traits might contribute to predicting IGT decision group despite lack of significance in the statistical tests. To investigate this hypothesis, a set of classifier models were used to evaluate the degree to which the trait measures might contribute to predicting IGT Group as compared to the WCST, Digit Span, CET and Lifestyle measures. Given the significant association found between performance on the WCST and IGT Group, the WCST was expected to perform better than chance in predicting group membership. If there were regularities in the association between scores on the other measures and IGT Group, these regularities should be exploitable by the classifier model and should contribute to predicting group membership above chance.

To predict the membership of each participant in one of the three IGT decision groups, multinomial logistic regression models were fit to each of the sets of predictors (the traits, WCST, Digit Span, CET, and lifestyle measures), and the classification accuracy was computed using the cross-validation procedure discussed in the Methods section of this chapter. The results of this analysis are shown in Figure 5.5 which shows the cross-validated prediction accuracy for each of the sets of predictors that were tested. The dotted lines in the figure are reference levels of accuracy corresponding to chance (randomly assigning each participant to one of the three clusters) and to a null model (which always assigns participants to the modal group, in this case the Frequency-Sensitive group that contained 39% of the participants). As expected, when the single-factor scores on the WCST were used as predictors, they contributed to prediction accuracy, yielding a mean accuracy of 0.444. In contrast, the lifestyle measures (Life), trait factor scores using both the four- and twelve-factor solutions (T4f, T12f), and the CET contributed nothing to prediction over and above the null model. A model using scores

on the Digit Span (D) as predictors yielded an accuracy of 0.407, contributing little relative to the null model.

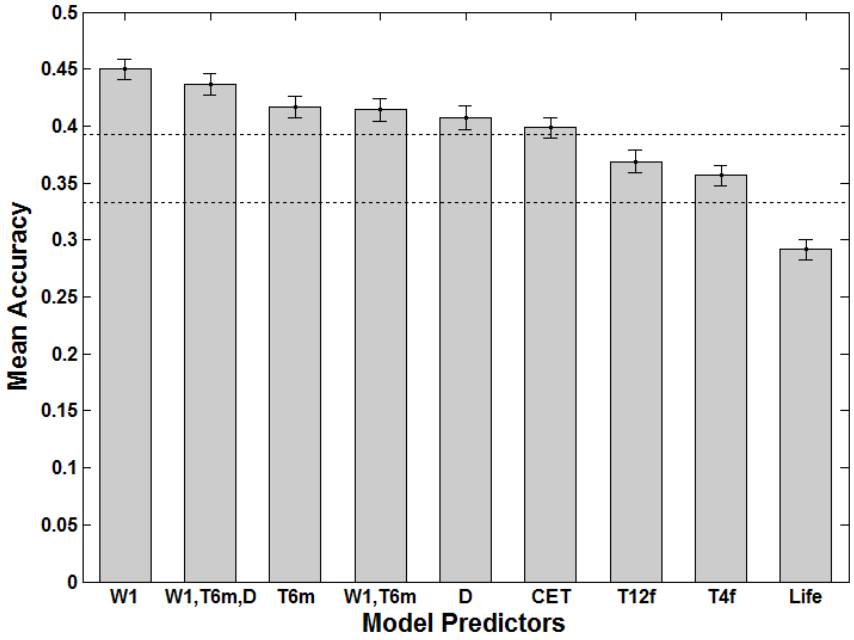

Figure 5.5 Mean cross-validated accuracy in predicting the membership of each participant in one of the three IGT decision groups. Each bar represents a model based on a set of unique predictors: the five lifestyle measures (Life), the four-factor trait scores (T4f), the 12-factor trait scores (T12f), the seven WCST measures (W7), the Cognitive Estimate Test scores (CET), the Digit Span scores (D), six individual trait scores found to be closest to significance in the univariate post-hoc analysis (T6m), the single-factor WCST scores (W1), and two models that combine the WCST single factor trait scores with the trait measures (W1,T6m) and with the trait measures as well as the Digit Span (W1,T6,D). Error bars are standard errors of the mean classification accuracy as computed by the cross-validation procedure.

As discussed in the last section, it was possible that a small set of the individual trait measures might contribute to prediction accuracy, despite being insignificant or marginally significant. To test this hypothesis, the six traits found to have p-values less than 0.15 in the exploratory tests (c.f., Figure 5.4) were tested as predictors alone (model T6m) as well as combined with the WCST factor scores (model W1,T6m) and with the WCST factor scores and Digit Span scores (model W1,Tgm,D). These six measures were Dysfunctional Impulsivity, Financial Risk Taking, Perseverance, Order/Regularity, Regret, and the Beck Depression Index. Interestingly, these six traits independently yielded a mean accuracy of 0.417 as compared to accuracy of 0.444 obtained from the WCST factor scores. When combined with the WCST predictor, however, these traits did

not contribute additional accuracy to the prediction model (c.f., Figure 5.5). The contribution of each of the six traits was then evaluated by examining the exponentiated β coefficients (e^{β}) in the logistic regression equations (Table 5.8). The reason for interpreting e^{β} values rather than the β coefficients is that they indicate the increase/decrease in the log-odds of the outcome group relative to the reference group for each unit increase in the associated predictors. Of the six trait measures included in the model, only Order/Regularity was significant in the prediction equation for the Frequency-Sensitive relative to the EV-LowFreq group, and the e^{β} value indicated that a participant one unit higher in this compulsiveness measure was 1.7 times more likely to be in the Frequency-Sensitive group than in the Low-Frequency group. The Financial Risk Taking and Beck Depression Index approached significance for this same equation, and their respective e^{β} values indicated that participants in the Frequency-Sensitive group tended to have higher scores in self-reported depressiveness and lower scores in risk taking. No coefficients were within, or near, significance for the prediction equation contrasting the EV-HighFreq group with the EV-LowFreq group. Taken together with the other analysis of the trait measures, these results again find little evidence for a robust association between IGT group and the studied traits.

Table 5.8 The e^{β} coefficients in the six-trait logistic regression model.

Measure	Frequency - Sensitive vs. EV-LowFreq	EV-HighFreq vs. EV-LowFreq
Order & Regularity	1.70**	1.36
Beck Depression Index	1.56*	1.18
Financial Risk Taking	0.66*	0.66
Regret	1.31	0.89
Dysfunctional Impulsiveness	0.82	1.33
Perseverance	1.07	0.71

Note. *=p<0.10, **=p<0.05. Values in the table are exponentiated β coefficients, which indicate the increase or decrease in the log-odds of the outcome group relative to the reference group for a unit increase in a predictor.

Discussion

In this study, I sought to characterize three previously identified group differences in decision making style in the IGT. These groups were shown to differ in their preferences over the four decks. The Frequency-Sensitive group preferred the two decks with the

common attribute of lower loss frequency relative to the other decks. The EV-LowFreq and EV-HighFreq groups showed a common preference for options with higher expected value, but differed in their preference for loss frequency. The primary aim of this study was to identify, aside from the demonstrated differences in performance in the IGT, other important ways in which the three groups might differ.

The approach pursued in addressing this aim was shaped in part by an existing body of research that has sought to characterize differences in IGT performance in terms of trait and affective correlates. Trait-based research on the IGT has been largely motivated by the fact that certain clinical populations that tend to have extreme scores on various trait measures also tend to perform disadvantageously on the IGT. The premise that these same traits are also robust correlates of IGT performance has not yet been established: findings to date have found only weak associations and these results have not been consistently replicated. In this study, I hypothesized that by grouping participants according to the three decision styles, I might be able to reveal more robust trait correlates of IGT performance. Included in the study were a large set of candidate measures (assessing a smaller set of constructs) chosen because they had been previously studied in the IGT literature or found to be associated with decision making, more broadly defined.

The approach taken in this study was also shaped by a parallel body of research that has attempted to disentangle what appears to be a complex relationship between the IGT and executive function. There are two views on what processes underlie performance on the IGT, each based primarily on neuropsychological evidence. One view is that the IGT is, foremost, a task driven by affective processes that serve to constrain and guide decision making. Proponents of this "affective view" have pursued evidence that might show dissociations between IGT performance and cognitive functions typically associated with "cold" aspects of decision making (Fellows, 2007). An alternative view suggests that advantageous performance in the IGT may be more a function of executive processes such as reversal learning and set-switching, and of cognitive flexibility. Proponents of this more "cognitive view" of the IGT have sought evidence showing that IGT performance and more "cold" executive functions are integrally related (Grant & Berg, 1948; Spreen & Strauss, 1998b). Much work is left to be done to reconcile these two views, and in this study I hoped that by investigating differences in IGT decision group

using candidate cognitive correlates concurrent with candidate trait correlates, I might provide new evidence to help do so.

After demonstrating that differences in demographic and self-reported lifestyle measures were not significantly associated with differences in IGT decision style, I utilized a range of approaches to try to reveal meaningful associations with a set of personality constructs putatively related to decision making in the IGT: (i) impulsivity, (ii) the closely related traits of compulsivity, indecision, negative perfectionism and maximizing behavior, (iii) behavioral drive and goal-seeking, (iv) risk-taking and sensation seeking behavior, and (v) general affect and temperament. I found no evidence for strong associations between these traits and differences in performance on the IGT, despite analyzing the traits at the level of both individual scales as well as factor-analyzed composites. Exploratory post-hoc analyses and prediction modeling offered some evidence that a small set of individual measures may be associated with frequency-sensitive behavior, but given the exploratory nature of these analyses, these associations should be viewed with caution. Of primary significance was the lack of evidence found to support the well-accepted claim that disadvantageous performance on the IGT is closely related to impulsivity. I included in this study three differing and widely used assessments of impulsive behavior as well as subscales from related assessments that also measure facets of impulsivity. These measures were analyzed individually as well as via the construct of impulsivity that emerged as one of the factors in the factor analysis. As a factor, impulsivity was not significantly associated with IGT decision making; of the individual impulsivity measures, only one was found significant in univariate post-hoc comparisons (the Dysfunctional Impulsivity scale of the DII), but this measure did not contribute significantly to predicting decision style in the IGT.

Concurrent with the trait analysis, I sought to identify associations between IGT decisions style and performance on three cognitive tasks (the WCST, the Digit Span, and the CET) as well as with a self-reported measure of mathematical ability. I found evidence for a significant association between IGT group and performance on the WCST, but not for any of the other measures. In particular, poorer performance in sorting cards on the WCST was associated with the group of participants found to be more sensitive to the frequency of losses as compared to the expected value of the IGT decks. This suggests that sensitivity to expected value – an important quantity on which this group

differed from the other two -- may be tied to executive functions tasked by the WCST. While the underlying processes tasked by the WCST are debated, it is thought to engage multiple components of executive function including categorization, rule learning and maintenance, set-shifting/reversal learning, and response inhibition (Fellows & Farah, 2003, 2005). The analysis of individual WCST measures revealed that poorer WCST performance by the Frequency-Sensitive group was significantly associated with perseverative- rather than non-perseverative errors (together, the two type of errors that can be made on the task). Perseverative errors occur when, faced with a change in the sorting rule, participants continue sorting cards based on the last rule in force. Participants in the Frequency-Sensitive group made more perseverative errors, and on average persisted in perseverative responses for more trials than participants in the two EV-Sensitive groups. These participants, however, did not seem to take longer to learn a sorting rule (there was no significant difference in trials-to-first-category across the three groups), which suggests that their perseverative errors were more likely due to poor performance in cognitive flexibility and set-shifting (reversal learning) than to categorization and learning abilities.

Interestingly, one prominent account of decision making in the IGT – associated with the cognitive view – is that disadvantageous decision making is associated with reversal learning (Bechara & Damasio, 2005; Bechara, Damasio, et al., 2000). In both the affective and cognitive conceptualizations of the task, it has been accepted that (i) the bad decks appear good in the first trials of the game, and therefore (ii) to perform advantageously, participants must learn that the good decks offer better longer-term payoffs and forego choosing from the bad decks in favor of the good decks. The two views differ in their claims about what processes underlie this transition from the bad decks to the good decks. Proponents of the more affective view have proposed that successful transitions are guided by affective signals originating in the body (or cortical representations of the body) and that are the result of learned associations ("somatic markers") that signal the value of the decks based on experienced payoffs (Bechara, et al., 1997; Bechara, et al., 1996). Patients with damage to vmPFC perform the IGT (and real life decision making tasks) disadvantageously, and because the vmPFC is thought to play a role in the integration of affective bodily signals with cognition, this involvement of this brain area in IGT performance is associated with the affective view of the task. By this view,

disadvantageous performance is thought to be related to a reduced (or impaired) ability to integrate delayed reward values into the learned somatic markers ("myopia for the future"). Behaviorally, this condition is thought to manifest itself as impulsivity. The primary empirical evidence supporting the specific association between vmPFC and affective signaling (as opposed to other possible component processes of decision making) comes from studies that have measured skin-conductive responses (SCRs) and demonstrated that that advantageous participants in the IGT develop anticipatory responses prior to selecting cards from the bad decks, while disadvantageous patients fail to show such responses (for a review, see B. D. Dunn, et al., 2006). Although these findings have been independently replicated, the link between anticipatory SCRs and affective signaling is debated (Fellows, 2007; Fellows & Farah, 2003, 2005). Farah and Fellows have advanced a somewhat different account, arguing that the transition across decks is subserved by processes involved in reversal learning (Fellows & Farah, 2005). The vmPFC is known to be involved in reversal learning, and by manipulating the payoff schedule in the decks, Farah and Fellows have shown that IGT performance in vmPFC patients improves when losses are experienced in the early trials and no reversal from the bad decks to the good decks is required (Jameson, et al., 2004). Farah and Fellows have also shown that patients with damage to dorsolateral prefrontal cortex (dlPFC) are impaired in the IGT, but critically these patients showed no improvement when the deck reversal requirement was eliminated. These neuropsychological studies of the IGT suggest that while affective signaling (and the putative traits associated with it) may guide behavior in the IGT, performance on the task is likely to be more broadly related to executive processes that subserve problem solving aspects of decision making in the task. The results of the present study in healthy participants are certainly consistent with this view. Among the large set of measures studied, only poor performance in terms of perseveration and set-shifting in the WCST emerged as correlates of disadvantageous performance.

The results of the present study are also consistent with evidence from related work that studied the effects of dual-tasking on performance in the IGT (for reviews see: Anderson, 1991; Ashby & Maddox, 2005; Bruner, Goodnow, & Austin, 1956; Medin & Smith, 1984; Murphy, 2002; Murphy & Medin, 1985; Osherson, Wilkie, Smith, Lopez, & Shafir, 1990). While performing the IGT, participants periodically engaged in secondary

tasks: either the maintenance and manipulation of a list of digits, or articulatory suppression. The secondary digit task was found to disrupt performance on the IGT as well as skin-conductance responses; articulatory suppression had no significant effect on either measure. Similarly, in the present study I found that while performance on the WCST was associated with the more disadvantageous IGT group, performance on the Digit Span was not. Taken together, the findings of in this study and the dual-task study suggest that differences in IGT performance are related to the more complex components of executive function (for example rule manipulation and set-shifting) than to lower-level processes such as verbal buffering and rehearsal.

One interesting attempt to reconcile the affective and cognitive views of the IGT has been put forth by Brand and colleagues. They conducted a study comparing performance on the IGT, the WCST and the Game of Dice Task (GDT) – a decision task in which participants are given explicit knowledge of payoff probabilities. The study found that performance on the WCST and the GDT was correlated more strongly with performance in the later trials of the IGT, than in the earlier trials. The authors argue that the early trials of the IGT involve decision making under ambiguity, while the middle and later trials involve decision making under risk. In the early trials, participants have little knowledge of the payoffs contained in the decks and must learn them through experience. As participants accumulate experience with the payoffs, their knowledge becomes more explicit, fundamentally changing the nature of the task. This view of the IGT is consistent with the findings from the clustering study reported in this dissertation. In this study I found that in each of the three groups, the participants' final pattern of preferences tended to emerge in the second or third block and persist for the duration of the task. That their preferences were stable across the middle and later blocks of the task suggests that participants were not engaged in reducing ambiguity through trial-and-error learning in these later blocks, but rather that they had developed explicit preferences and were involved in decision making to apply these preferences in the probabilistic context of the task.

The primary aims of this study were to further test the associations between IGT performance and traits previously studied in the literature, and to investigate the relative efficacy of trait and cognitive measures in characterizing differences in the IGT. These two aims were accomplished. In finding (i) a significant association between IGT decision

style and performance in the WCST, and (ii) no significant associations with the Digit Span or with trait measures previously implicated in IGT performance, the current study provides evidence suggesting that differences in the IGT are more likely due to cognitive processes involved in executive control than to processes involved in affective signaling (and the putative traits associated with these processes). However, given the large number of candidate measures investigated, the current study provides more information about the factors that do not help characterize IGT decision style than those that do. Participants in the Frequency-Sensitive group perform the IGT more disadvantageously and this difference in their performance was associated with differences in the components of executive function tasked by the WCST. However, the current study revealed little else about how the Frequency-Sensitive group differs from the two EV-sensitive groups. The study also revealed no evidence to help characterize differences between the two EV-sensitive groups. Further characterizing differences in the three identified IGT decision groups is necessary to provide a more complete conceptualization of individual differences in this task. The results presented in this chapter suggest that studies focused on cognitive factors involved in problem solving and executive control are likely to be a productive agenda to pursue.

CHAPTER VI. GENERAL DISCUSSION

Summary of Results

Decision making behavior in the IGT has almost exclusively been characterized based on population-level analysis using a single univariate measure of performance that assumes participants perform the task based on the relative expected values of the four decks. While it is certainly possible that (i) IGT performance is well represented by population-averaged patterns of performance, (ii) that individual differences are well-captured by quantitative differences in univariate measures, and (iii) that the most appropriate measure is expected value (e.g., *%Good* or *%EEV*), in the present work I sought to test the strength of these assumptions and in doing so also challenge the validity of the current conceptualization of the task.

I used computational models to determine how well decision making in the IGT is accounted for by single model that endogenously computes expected values based on lower-level procedural learning processes. The choice of model was not based on attempts to directly model the core features of the task, but rather was independently motivated by a known connection between performance on the IGT, damage to brain areas involved in reward-based learning, and a class of computational models known to reproduce observed neural data in these brain areas. I found that the class of reinforcement models, of which the widely accepted expectancy-valence model is a member, can account for the core phenomena in the task, when considered in the aggregate. Critically, I found that the addition of a frequency-avoidance term to the definition of reward yielded a model that provided a better overall account of aggregate performance than the standard model. However, considered at the level of individuals, this better model was the best fit for only a minority of participants. A range of variant models each were able to better capture performance for small subsets of participants than the best model. A modal model of the task therefore remains elusive. The fact that the RL class of models was unable to capture the performance of a large percentage of participants suggests there are forces at work that may be beyond the reach of a

theoretical framework that includes only lower-level reward-based learning mechanisms. This dissertation focused on the RL class of models motivated by their close relation to the Somatic Marker Hypothesis as well as by the goal of investigating this class of models more fully than has been done in prior computational work. It is important to acknowledge, however, that in limiting the computational study to RL models, higher-level declarative models have been ignored. The results of this dissertation suggest that declarative decision processes may have a greater influence on behavior in the IGT than has been suggested by the Somatic Marker Hypothesis and by the prevailing conceptualization of the task. Taken together, the finding that the RL models were a poor fit for a large minority of participants, the failure to find significant associations between IGT decision making and trait measures, and the finding of a significant association between the IGT and WCST provide converging evidence that investigating the IGT from a more declarative perspective may be a fruitful avenue for future work; this future work might make contact with the large and well-developed literature on concept formation and hypothesis testing (For reviews, see: Anderson, 1991; Ashby & Maddox, 2005; Lamberts & Shanks, 1997; Medin & Smith, 1984; Murphy, 2002; Murphy & Medin, 1985; Osherson, et al., 1990). An additional challenge to the application of the RL framework to modeling the IGT comes from a recent neuroimaging study demonstrating that exploratory behavior may be supported by a different neural substrate than the type of exploitative behavior associated with choices based on learned value estimates (Daw, et al., 2006). The authors of this study suggest that exploratory behavior may be due to frontal control mechanisms that override exploitative behavior. Although initial exploratory behavior in the RL framework is endogenously generated as result of setting value estimates to the same initial value, exploration and exploitation are instantiated in a unitary mechanism (the choice function) and as such are not consistent with the neural data. Explicitly modeling exploratory behavior as an independent control mechanism may also be a fruitful area for future research.

Motivated by the results of the computational study, I then used a multivariate clustering procedure to determine more robustly whether decision makers in the IGT are better characterized by a single pattern of performance differing quantitatively in their sensitivity to expected value, or by multiple patterns of performance based on decision attributes other than, or in addition to, expected value. Consistent with the results of the

computational study, the clustering results revealed three subsets of participants differing qualitatively in the ordering of their preferences over the four decks – three different decision styles. I found that in the first block of trials, all participants performed similarly, exploring each of the four decks with a notable preference for deck B. After this first block, the three decision styles began to emerge and each persisted for the remainder of the task. Participants exhibiting the EV-LowFreq style showed a preference for a deck with positive expected value and low frequency losses (deck D). The participants in the EV-HighFreq cluster also showed a preference for a deck delivering positive expected value, but showed a concomitant preference for losses that occurred with high frequency. Tellingly, when participants associated with each of these two decision styles depleted the cards in their preferred deck, they shifted their selections to the other deck offering positive expected value, suggesting strongly that their choices were based foremost on sensitivity to expected value, with loss frequency a secondary attribute. I also found a third pattern of performance associated with participants who preferred the two decks that delivered low frequency losses (decks B and D).

I interpret the performance of participants who showed a combined preference for decks B and D as being driven by sensitivity to loss frequency, and refer to this decision style as Frequency-Sensitive, but this interpretation deserves further justification as it is possible that one or more other confounded attributes might equally well characterize decks B and D. The primary features of the task are the gain amounts, loss amounts, and net payoffs[12] that participants experience as they select cards. Plausible attributes that might influence choice behavior are the expected values, variances, and frequencies of these quantities as summarized in Table 6.1 for each of the four decks. What our results demonstrated was that participants in the putative Frequency-Sensitive cluster showed a combined preference for decks B and D over decks A and C, and over a preference for any one deck. Inspection of Table 6.1 reveals that loss frequency is the only attribute

[12] Net payoff in the task is the gain amount on a winning trial, and the sum of the gain amount and loss amount on a loss trial. Net payoffs are not explicitly represented in the task, but it is plausible that a mental representation of this quantity is available to participants either via implicit or explicit processes. Maia and McClelland (2004) provide evidence that net payoffs and expected value quantities are available to participants. Whether these quantities actually influence choice behavior has not been fully determined.

common to decks B and D and not shared by decks A and C. While there may be other more complex attributes of choice that I have omitted in this analysis, parsimony supports the conclusion that loss frequency is the appropriate attribute for characterizing the participants that demonstrated a preference for decks B and D. This finding is consistent with another recent study that found frequency was the dominant attribute of choice in children age 6-15 and that frequency and magnitude attributes were utilized in young adults aged 18-25 (B. D. Dunn, et al., 2006).

Table 6.1 Decision attributes associated with decks in the IGT.

Attribute	A	B	C	D
EV[a] of Gains	High ($131)	High ($131)	Low ($66)	Low ($66)
EV[a] of Losses	High (-$203)	High (-$203)	Low (-$33)	Low (-$33)
EV[a] of Net Payoffs	Negative (-$72)	Negative (-$72)	Positive (+$33)	Positive (+$33)
Std[b] of Gains	High ($23)	High ($23)	Low ($14)	Low ($14)
Std[b] of Losses	Medium ($116)	Very High ($630)	Very Low ($24)	Medium ($100)
Std[b] of Net Payoffs	Medium ($109)	Very High ($625)	Very Low ($25)	Medium ($100)
Loss Frequency	High (50%)	Low (10%)	High (50%)	Low (10%)

Notes. [a] EV denotes expected value. Expected values were computed across the set of cards contained in each deck. [b] The text refers to variance, but the table shows standard deviations (std) to allow comparison of the values in terms of dollar amounts rather than squared dollar amounts.

I cautiously interpret the behavior of participants who showed a preference for either deck C or D as being sensitive to expected value and I have therefore labeled their assigned clusters as EV-LowFreq and EV-HighFreq depending on whether they preferred the low or high frequency deck. The fact that the design of payoffs in the IGT confound several attributes has been previously raised in the literature, for example (B. D. Dunn, et al., 2006). This issue is evident from the attribute comparisons shown in Table 6.1. Advantageous participants prefer decks C and D, and the clustering results show that they also tended to prefer one of the two decks, but not both. What attributes are common to decks C and D, and not shared with decks A and B? There are at least four possibilities. Decks C and D have in common that they both deliver low average gains. It is not likely that subjects prefer these decks *because* they offer average gain amounts, so I rule out this attribute. Also common to these two decks is the fact that the average magnitude of the losses in these decks is lower than in decks A and B. As noted

by Dunn (Kahneman & Tversky, 2000), advantageous performance could be an outcome of behavior that ignores gain magnitudes and instead seeks lower magnitude losses. The expected value of losses is therefore a quantity that may plausibly drive the behavior of participants in the EV-LowFreq and EV-HighFreq clusters. Because the expected value of net payoffs is highly correlated with the expected values of the losses, it too is a plausible basis of performance. The fourth attribute common to decks C and D is that the gains delivered by these decks are lower in variance than the gains in decks A and B. It is therefore also possible that advantageous performance is driven by sensitivity to the variability in the magnitude of gains. The variance in losses and net payoffs are also shared by decks C and D, but also by deck A. The variance differences between decks D and A are very small and it unknown whether such small differences are sufficient to influence performance in the task. In summary, while I think that it is unlikely that participant preferences for either deck C or D are influenced predominantly by attributes of variance rather than expected value, I leave open this possibility for determination by future work. The set of attributes outlined in Table 6.1 (and possibly others as well) need to be independently manipulated and tested to determine which of them influence advantageous behavior. Critically, I suggest such tests should be done with the unit of analysis being groups (grouped based on the three decision styles identified in this study) rather than populations.

Linking Modeling and Clustering Results

In the computational study, patterns of deck choices were generated endogenously by the mechanisms instantiated in the set of models that were considered. In the clustering study, patterns of empirically observed deck choices were analyzed to identify dimensions on which participants differed in their decision behavior. The results of these two studies independently suggested that loss frequency is an important attribute associated with decision making in the IGT. The best-fitting model included loss frequency as a component of the reward function, and loss frequency was shown to be an important attribute associated with differences in the three identified clusters. Given this convergence of results, I sought to further investigate the association between the models and the three identified clusters. Table 6.2 gives a distribution of the best fitting models for participants in each of the three clusters.

Table 6.2 Distribution of best fitting models within clusters.

Cluster	N	Base	Simple Average	Decay	Pursuit	Reinforcement Comparison	Risk Models**
EV/HighFreq	16	6%	13%	13%	0%	19%	50%
EV/LowFreq	11	27% /0%*	18%	0%	18%	0%	36% /63%*
Freq-Sensitive	12	17%	17%	0%	17%	0%	50%

* The second value gives the percentage of participants without the complexity correction imposed by the use of the BIC criterion. Because the Base model was nested within the Risk Sensitive models, comparison of these models without the complexity correction is appropriate. The 27% of the participants in the EV/LowFreq cluster that was best fit by the Base model were best fit by the Risk-Sensitive Loss Frequency model when the complexity correction was not considered. ** For clarity of presentation the six risk-based models were combined into a single column in the table.

It is evident from the table that for each of the three clusters, the set of Risk Models (which includes the best-fitting Risk-Sensitive Loss-Frequency model) fit more participants than any other model. This is the expected result if the modeling results and clustering results are meaningfully associated and serves to further corroborate the results of the two studies. I note, however, that this is not a necessary result. The risk-based models might have fit 100% of the participants in one cluster, a small number of participants in a second cluster, and no participants in a third cluster – an outcome that would have suggested the two studies were not capturing the same aspect of behavior. The pattern of the best-fitting models across the clusters is also suggestive of possible associations between model structures and the three clusters. For example, the Decay and Reinforcement Comparison models were the best fitting models only for participants in the EV/HighFreq cluster. These two models were both premised on lower fidelity of available information: the Decay model instantiating reduced fidelity in the maintenance of value estimates and the Reinforcement Comparison modeling learning as based on a single reference level of reward rather than separate reference levels for each deck. Taken together, this integration of the modeling and clustering offers a tentative suggestion that behavior by participants in the EV/HighFreq cluster may be driven by an inability to maintain information over time and/or possess decision processes that rely on reduced information sets. As a second example, the Pursuit model also shows an interesting distribution by cluster. This model fit a large percentage of participants in the two clusters that preferred lower frequency losses, but no

participants in the cluster that preferred higher frequency losses. The Pursuit model instantiated the idea that choice probabilities may be learned independently from value estimates, making frequently chosen options more likely to be chosen again. This result is well-aligned with the revealed association between the Frequency-Sensitive cluster and the WCST and further suggests that perseveration (switching behavior) may be associated with participants that exhibit a preference for avoiding losses.

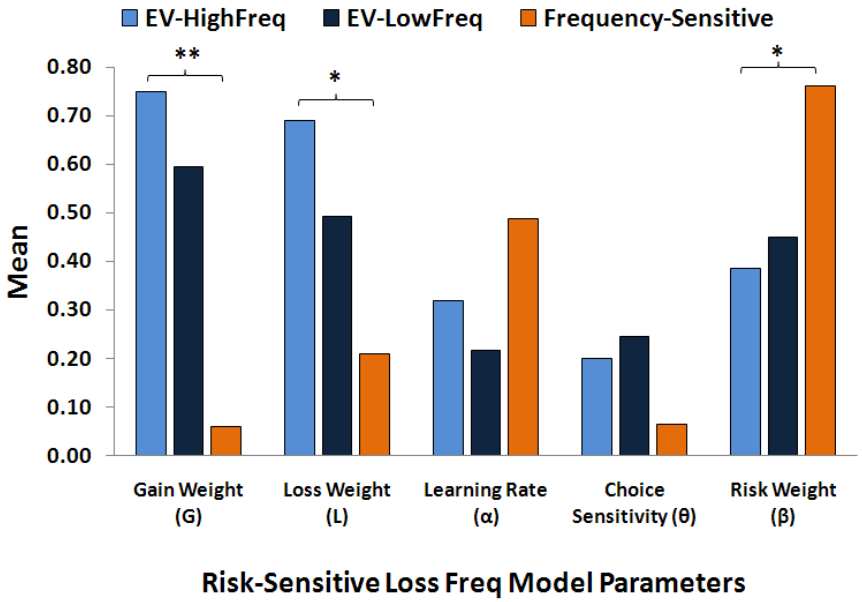

Figure 6.1 Association between model parameters and clusters. For participants in each cluster, the plot dhows the mean maximum-likelihood parameter estimates for each of the parameters in the Risk-Sensitive Loss-Frequency (RSLF) model. The estimated Beta parameter for participants in the Frequency-Sensitive was higher than for participants in the other clusters indicating that loss-frequency was weighted more heavily in fitting the model to these participants. The Gain and Loss weights were lower for participants in the Frequency-Sensitive cluster as compared to the other two clusters, further suggesting that choice behavior was associated more with loss-frequency than net payoffs for these participants.

In addition to considering the distribution of models across clusters, I also sought to corroborate and integrate the two studies by analyzing the association between the three clusters and the estimated parameters for the best-fitting Risk-Sensitive Loss Frequency model. This analysis is shown in Figure 6.1. If the Risk-Sensitive Loss Frequency model did in fact capture the same underlying constructs of expected value and risk that were captured by the clustering procedure, then there should be a meaningful association between estimated model parameters and three clusters. Specifically, based on the

results of the two studies one would predict that for participants in the Frequency-Sensitive cluster, the risk term should have been more heavily weighted (the β parameter in Equation 8c) than for participants in the other clusters. Likewise, the weighting of gains and losses (the G and L parameters in Equation 8c) should have been larger for participants in the EV/LowFreq and EV/HighFreq clusters as compared to participants in the Frequency-Sensitive cluster. Both of these predictions are supported by the results shown in Figure 6.1.

Reconceptualizing the IGT

Taken together, the computational modeling and clustering results revealed that the current conceptualization of the IGT is not complete, and together suggest an augmented conceptualization that is both complementary to the prevailing characterization of the IGT and that helps to further elaborate the current dimensional conceptualization of IGT performance based on the disadvantageous-advantageous continuum (Figure 6.1). The results of this dissertation clarify advantageous performance, by revealing two distinct subsets of advantageous participants that differ in their preferences for loss frequency. The results also help clarify disadvantageous performance by revealing a third decision style in which nearly equal proportions of disadvantageous and advantageous performers shared a preference for infrequent losses that seemed to prevail over their preference for expected value. This finding suggests disadvantageous performance might be better conceptualized in terms of high sensitivity to loss frequency combined with reduced sensitivity to expected value. This mapping between the widely accepted advantageous/disadvantageous conceptualization of performance and the three decision styles revealed in this study strongly suggests an alternative and multidimensional view on the nature of decision making in the IGT. The existence of the EV-LowFreq and EV-HighFreq clusters (together representing 61% of the 855 participants in the combined data set) underscores the well-established fact that decision making performance in the IGT is driven by sensitivity to expected value. The existence of the large Frequency-Sensitive cluster (representing 39% of the 855 participants) suggests that sensitivity to risk (the frequency of losses, and/or other possibly confounded measures of risk common to decks B and D not identified in this

study) is also an important driver of decision behavior. And critically, the fact that the two EV-sensitive clusters differ in sensitivity to frequency, and within the Frequency-Sensitive are two subsets differing in sensitivity to EV, suggests a general framework for characterizing individual differences in the IGT (Figure 6.2). Whatever the underlying mechanisms, decision makers in the IGT seem to differ dimensionally in their sensitivity to *two* attributes of choice: the expected value and the loss frequency associated with each of their decision options. This augmented conceptualization of the IGT is certainly consistent with well-established decision theory, and in this work I have identified and quantified these two attributes specifically for the IGT.

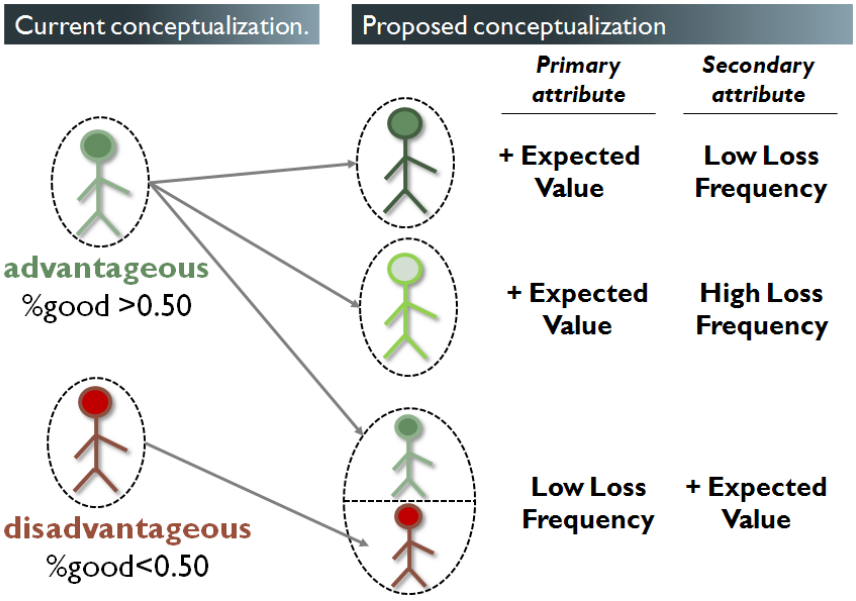

Figure 6.2 Current and proposed conceptualization of decision making behavior in the IGT. Participants currently characterized as advantageous share a sensitivity to expected value, but differ in their preference for loss frequency. Participants currently characterized as disadvantageous share a primary preference for low-frequency losses, and differ in their relative sensitivity to expected value.

How well does this two-attribute framework capture individual differences as revealed by the 20-feature clustering procedure? Figure 6.3 depicts each of the 855 participants from the combined data set in terms of their location in two-attribute decision space, highlighting the participants based on their assignment to the three decision styles. I measured the attribute "sensitivity to expected value" (horizontal axis) as the percentage of selections made from the two decks that deliver positive expected

value (C and D), and the attribute "sensitivity to loss frequency" (vertical axis) as the percentage of selections made from the two low frequency decks (B and D). I computed these two measures across the final three blocks of the task to better reflect participants' stable preferences after the initial period of learning. To check that the two measures did in fact capture two different attributes of choice, I computed the Pearson correlation coefficient and found a small ($r=-0.243$, $r^2=0.059$; $p<0.0001$) inverse relationship between the two measures thus supporting the idea that these measures do capture two different constructs. Figure 6.2 helps to clarify the relationship between the standard characterization of the IGT in terms of *%Good* (the vertical line at 0.5) and the proposed framework in which participants are *also* concurrently characterized in terms of their sensitivity to frequency of losses. Viewed in this two-attribute space, the three decision styles are well defined: the EV-LowFreq (dark blue) and EV-HighFreq (light blue) clusters occupy the region of high sensitivity to expected value and participants in these clusters are distinguished by differences in their sensitivities to frequency; the Frequency-Sensitive (orange) cluster occupies the upper frequency-avoiding region, and participants in this cluster differ in their relative sensitivities to expected value. The region in two-attribute decision space corresponding to low sensitivity to expected value combined with frequency seeking behavior (the lower left region) is unoccupied indicating that no participants exhibited a decision style demonstrating a preference for deck A or for a combination of decks A and C. Deck A is inferior on both attributes of choice (*c.f.* Table 6.1), and as a result it is likely the case that its inferiority is highly salient making it easy for all participants to avoid regardless of their individual differences in sensitivity to the two decision attributes. Self-reports obtained by participants during a pilot study support the claim that the inferiority of deck A is salient to participants, and this is certainly borne out by the empty region in lower-left corner of Figure 6.3.

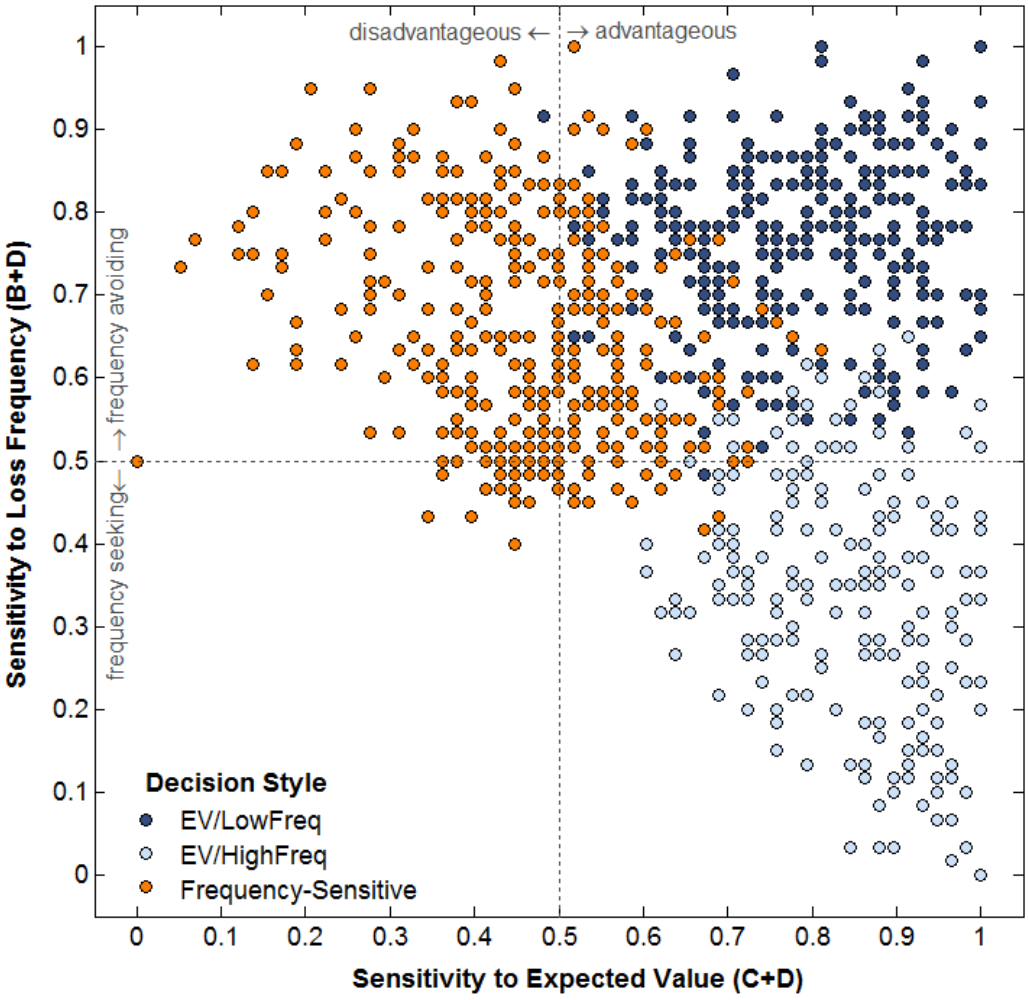

Figure 6.3 Two-attribute framework for conceptualizing individual differences in the IGT. Performance is conceptualized as being driven by differing levels of sensitivity to expected value and sensitivity to loss frequency. These two measures are quantified as the proportion of cards selected from decks C and D, and decks B and D, respectively. Shown are the 855 participants (points on the plot) from the combined data sets plotted in the two-attribute space based on their performance. The decision style assigned to each participant by the clustering procedure is represented by the color of each point in the plot. The plot therefore represents a mapping of participants from the 20-dimensional space in which they were clustered to 2-dimensional space represented by the two attributes.

The two-attribute framework seems to perform well in capturing the three decision styles identified in the clustering study. However, it is important to note that it ignores all temporal aspects of the task (i.e., it maps the 20-dimensional temporal pattern of performance into an atemporal 2-dimensional space). The three decision styles themselves are also largely uninfluenced by the temporal features of performance and

instead are distinguished in terms of relatively static preferences following the first block of trials. Temporal aspects of the task do play a role, and this is evident in Figure 6.3 by the presence of Frequency-Sensitive participants in the EV-Low region and by EV-Low participants in the region occupied by the Frequency-Sensitive participants: these "outliers" were found more similar to the other members of their cluster than to members of other clusters, which indicates that there are aspects of their four-deck-by-five-block patterns of performance not captured in the two-attribute framework. Also, the patterns we identified in the six-cluster solution suggest that within the three decision styles, participants may differ more subtly in the temporal aspects of their performance. However, the data clearly selected the three-cluster solution over the six-cluster solution. Whether temporal differences in performance are artifacts of the data set used in this study (e.g., picking up on depletion of cards from two of the decks), or indicative of subtypes of decision style that generalize to other data sets remains to be determined.

One important set of questions left unanswered in the current study is what fundamental factors underlie the differences in performance across the three clusters. What factors lead participants in the Frequency-Sensitive and EV-LowFreq groups to prefer the decks with lower loss frequency, and the participants in the EV-HighFreq group to prefer higher loss frequency? What factors lead decision makers in the Frequency-Sensitive group to be seemingly less driven by expected values than decision makers in the other two groups? The exploratory studies reported in Chapter V did not provide many answers to these questions. Among the many demographic, trait and cognitive measures studied, only an association between IGT decision group and executive function as measured by the Wisconsin Cart Sorting Task was found. In particular, decision makers in the Frequency-Sensitive group were found to perform poorly as compared to decision makers in the other two groups. Contrary to the often suggested (but largely unproven) claim that disadvantageous performance on the IGT is associated with impulsivity and risk-taking, the results of studies in Chapter V suggest that if traits have explanatory value in understanding differences in IGT decision making, their contribution is likely quite subtle and high-power studies will be needed to robustly identify and elucidate their effects. Overall, the studies of Chapter V provided more evidence for what measures do *not* help explain differences in performance on the

IGT than those that *do*. While fundamental differences underlying the three identified decision groups remain to be found, there are several possible theoretical accounts that might be sorted out in future work.

One possibility is that frequency serves as a measure of risk, and that the clusters of participants differ in their risk preferences. Loss aversion is a well-established finding in behavioral economics (Kahneman & Tversky, 2000) and it is possible the identified differences in behavior in the IGT reflect differences in this aspect of risk – either differences in aversion to loss occurrence or differences in response to losing current wealth (an endowment effect). According to this risk-based hypothesis, for participants in the Frequency-Sensitive cluster, the expected value of the decks becomes secondary to the more rewarding outcome of avoiding loss events; for participants in the EV-LowFreq cluster, expected value becomes primary, but behavior is also guided by avoiding losses and therefore these participants prefer deck D over deck C.

Another hypothesis (not necessarily mutually exclusive) is that loss-frequency serves as a proxy for ambiguity and that group differences are due in part to differences in tolerance for ambiguity. Ambiguity aversion (separate from loss aversion) has also been demonstrated experimentally to be an attribute of choice (Ellsberg, 1961), with the famous Ellsberg Paradox being one early example (Shurman, Horan, & Ntuechterlein, 2005). By this account, the high-frequency decks (A and C) provide more information about the magnitude and frequency of losses. For example, in the first 20 draws from deck A, a participant would experience 20 gains of varying amounts, and 11 almost evenly spaced losses in the range (-$150,-$350) which provides a consistent experience with the losses; likewise, the first 20 draws from deck C yield 20 gains of varying amounts and 11 losses in the (-$25,-$75). In contrast, in the first 20 draws from deck B, a participant would experience one loss in the first nine trials (representing an experienced frequency of about 11%), and a second loss five trials later (representing a point frequency of 20% for this loss); deck D yields a similar experience as deck B. The two decks with high frequency losses (A and C) therefore provide information about their payoffs more rapidly and more consistently, and it is possible that participants more averse to ambiguity prefer these decks. Although this ambiguity- or information-based account would seem to predict that some participants should prefer both decks A and C,

participants seem to learn quickly that the payoffs in deck C are inferior and therefore no participants show a stable preference for this deck.

A third possible account of group differences is that they are due to differences in executive function or problem solving style and/or capability. Loss frequency may be a more salient attribute of the task than other attributes, for example it may be easier to track differences in the frequency of occurrence of losses across the decks than to track the average magnitudes of the losses relative to the average magnitudes of the gains as would be required to process expected value. According to this hypothesis, the frequency-sensitive participants may be pursuing a lower-effort approach to decision making (easier to avoid losses than track expected values), either because they are "lazy" or less-motivated or because they are cognitively limited relative to other participants in terms of executive function and/or memory capacity. If participants in the Frequency-Sensitive cluster had more limited executive function, the task would be in effect more demanding for them and this might result in the pursuit of a simpler approach to problem solving. In some ways this lower-effort problem solving approach of avoiding losses is akin to a "Maxi-Min" strategy whereby decision makers pursue options that seem to minimize outcomes defined as worst-case. A variant of this account is that the limitation in executive function is due specifically to lower than average performance in functions associated with set-shifting and cognitive flexibility. The revealed association between IGT decision group and WCST performance offers some support for this hypothesis, as do results of the reversal learning studies in vmPFC patients that were discussed in Chapter V.

Contributions and Conclusions

The primary contributions of this dissertation research are that it revealed important limitations in the current conceptualization of the IGT and identified important ways in which this conceptualization could be made more complete. How might the proposed conceptualization come to bear on future performance analysis and inference using the IGT? Although univariate analysis of the IGT using *%Good* (total selections from decks C and D) is methodologically convenient, the proposed conceptualization suggests that a better account of performance may be achieved by adding loss frequency (total selections from decks B and D) as a second measure and using multivariate methods for

data analysis (c.f., Figure 6.3). Furthermore, the generalizability of the cluster prototypes demonstrated in Chapter IV suggests the possibility of using these prototypes as "group norms" for each decision style. These norms combined with a simple multivariate distance metric (Euclidean or Mahalanobis) could be used to classify participants for the purposes of group-wise univariate or multivariate inference (without having to perform the complex clustering procedure to do the classification). By doing tests for mean differences or tests of association using group as the level of analysis rather than population, the loss of power due to the smaller sample size in each group might be outweighed by the reduction in variability gained by separating participants into more homogeneous subsets than the population as a whole.

In terms of clinical implications, the results of this work suggest that a change of measure in assessment may be merited. Characterizing clinical populations based solely on sensitivity to expected value ignores the important second dimension of performance (loss frequency) that was revealed in Chapters II and III. Lack of uniformity in findings of studies testing for differences between clinical populations and healthy controls may be due in large part to the unitary focus on expected value (and perhaps also to differences across studies in the distribution of the healthy control among the three decision groups). For example, a recent study of schizophrenia patients found that the patterns of performance exhibited by these patients differed from both healthy controls and vmPFC patients, and these differences were associated with both payoff magnitudes and loss frequencies (Glimcher, 2008). How might the proposed conceptualization of the IGT be used to better characterize clinical populations? Rather than focusing only on whether a patient population does/doesn't perform more disadvantageously than healthy controls, the proposed conceptualization engenders a different set of questions (Figure 6.4). First, *is the patient population better represented by a single pattern of performance, or multiple patterns of performance*? This question can be addressed by applying the ensemble clustering procedure directly to patient decision data. Second, *does performance in a population of patients differ qualitatively or quantitatively from healthy controls*? If patients (considered either in aggregate or in identified groups) differ qualitatively from the three decision groups identified in this dissertation, this could be revealed by applying the clustering procedure to a combined data set and determining whether the patients emerge in an independent cluster, or co-associate in one or more of the clusters of

healthy participants. If patients differ only quantitatively, this could be manifested in either a difference in the distribution of patients across the three healthy decision groups (e.g., the patient group might be disproportionately represented in the Frequency-Sensitive cluster) or as a difference in location within the healthy decision groups (e.g., the patient group might populate an extreme region within the EV-HighFrequency cluster). Third, *what computational models provide a better fit to a patient population as compared to healthy controls*? While previous modeling work has sought to characterize patients in terms of differences in model parameters, the results of this dissertation suggest that characterization using multi-model inference may be more appropriate. By fitting a set of models to a population of patients and to healthy controls, inferences about differences in decision processes can be made by comparing differences in the structure of the best fitting models selected for each population.

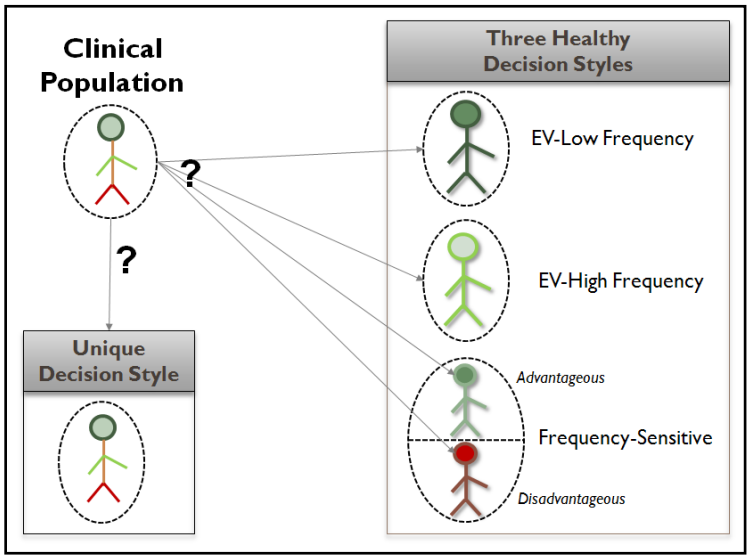

Figure 6.4 An alternative approach to characterizing impaired decision making in the IGT. The results of the dissertation suggest a clinical approach that seeks to conceptualize patient performance in terms of both expected value and loss frequency, and in terms of the three groups of healthy controls. A patient population may differ qualitatively and be best represented as a unique pattern of decision making. Alternatively, a patient population may be better represented as differing quantitatively from one or more of the patterns of performance identified in healthy controls.

In addition to revealing differences in decision style and suggesting a new framework for conceptualizing the IGT, the results of this dissertation offer some methodological contributions to the study of decision making. Multivariate clustering is a

useful approach for representing performance in a behavioral task perhaps more deeply than is typically amenable to univariate hypothesis-testing procedures. By characterizing performance in more detail, multivariate clustering methods can help reveal individual differences in performance and provide a way to test the validity of population-averaged measures prior to conducting univariate hypothesis tests. While clustering tools (e.g., k-means and hierarchical methods) are widely available and easy to use, the results of the ensemble clustering study offer a note of caution to the use of a single, off-the-shelf methods for clustering psychological data. As discussed in Chapter III, there is little guidance available to assist the modeler in choosing among clustering methods when comprehensive knowledge of the underlying structure of the data is not available. In fitting individual clustering models to the IGT data using a range of clustering algorithms and distance metrics, I found that while some combinations of models and metrics yielded clustering solutions similar to those obtained with the more robust ensemble procedure, other combinations produced very different solutions that were often unstable across bootstrap samples. Interestingly, when I re-ran the ensemble clustering procedure on the data using only these ill-suited combinations of models and metrics, I recovered the same basic pattern of clustering solutions obtained with the full set of models and metrics. Taken together, these findings therefore emphasize the importance of diversification of method that is the hallmark of ensemble clustering procedures – and critically, these findings highlight the possibility of obtaining erroneous results from applying a single off-the-shelf clustering method to a data set in the absence of prior knowledge (e.g., number, shape, size, density and separability) of the natural clusters that might exist. Ensemble methods are widely used to uncover underlying patterns in gene-expression data, and here I have shown how these methods might be put to use in the analysis of psychological data.

Future Directions

Although the results of this dissertation improve on the current computational model and theoretical conceptualization of decision making in the IGT, it is important to note that the augmented model and conceptualization identified in this dissertation are still incomplete in many ways that suggest the need for further research. First, the lower-level learning processes instantiated in the RL framework provide a computational

account of the core phenomena of the IGT, but it is possible (and in my opinion, likely) that a more complete mechanistic account will require that the RL model be extended to include mechanisms reflecting high-order cognitive processes. Behavioral economists have developed a range of models designed to explain normative departures in decision making under risk and uncertainty (Kahneman & Tversky, 2000; Maia & McClelland, 2004). Integrating these higher-level models with reward-based learning models may be a productive avenue for future research. Second, the three groups of decision makers revealed by the clustering study were shown to be robust across independent data sets, but how these groups differ in more fundamental ways remains largely unexplained. Traits that were plausibly relevant to decision making provided no account of these group differences, although this is an interesting result given the common claim that disadvantageous performance on the IGT is associated with impulsivity and risk-taking. The association between the IGT and the WCST identified in this study, taken together with the finding that the IGT *is* cognitively penetrable (Fischer, Corcoran, & Corcoran, 2007) suggest that further studies seeking associations between the IGT and cognitive measures of performance may be more fruitful than further work seeking to identify trait-based correlates. I proposed several cognitive-based accounts of how the three IGT decision styles might differ, and additional work linking IGT performance to differences in problem solving and executive function might be used to test competing predictions and to adjudicate among these accounts. Additional evidence to help further characterize the three groups might also be revealed by using neuroimaging to test for group differences in the patterns of neural activity evoked by performance of the IGT. Lastly, the IGT was designed for the purposes of clinical assessment rather than careful experimental manipulation, and as a result it is messy task and one in which many variables are confounded and outside of experimenter control. It is nevertheless an important task and as such, one that I felt was worthy of closer scrutiny. The IGT is, however, but one of many experimental decision tasks and extending the methods and findings of this dissertation beyond the IGT is an important next step and one that I hope to pursue.

Appendix A. Participant Screening Questionnaire

C&CN Laboratory -- Short Participant Questionnaire

Please respond to the questions below. This information is confidential and will be used by the Cognitive and Computational Neuroscience Laboratory only for the purposes of the experiment in which you have expressed interest. Our collection of the information below has been approved by the Institutional Review Board at the University of Michigan. Our methods of confidential and secure data storage have also been approved by this Board.

1.

Age:

2.

Have you had, or do you currently have, any of the following psychological/psychiatric conditions? (Please check ALL that apply)

- A ☐ ADD/ADHD
- B ☐ Alcoholism or Drug Dependency
- C ☐ Phobias (Clinically Diagnosed)
- D ☐ Obsessive Compulsive Disorder
- E ☐ Major Depression
- F ☐ Co-morbid Depression (Depression in the presence of a physical illness)
- G ☐ Bipolar Disorder
- H ☐ Schizophrenia/Schizophrenic Disorders
- I ☐ Anorexia/Bulimia
- J ☐ Post Traumatic Stress Disorder
- K ☐ Pathological Gambling
- L ☐ Personality Disorders (Borderline, Narcissistic, or other)

Other psychological or psychiatric disorder not listed above (please describe below):

- M ☐
- N ☐ I have never had, nor currently have any psychological or psychiatric disorder.

3.

Do you have a known history of head trauma (for example, a concussion) or neurological disease or damage (for example, a stroke, Alzheimer's disease)?

- A ○ No

Yes (if so, please describe head trauma, neurological disease or other damage below):

- B ○

4.

Are you currently enrolled in the course "Introduction to Psychology (Psych 111) at the University of Michigan?

- A ○ Yes
- B ○ No

5.

Are you a native/fluent English speaker?

- A ○ Yes
- B ○ No

6.

Are you currently a U-M employee?

- A ○ Yes
- B ○ No

7.

How did you hear about this experiment?

- A ○ Flyer/Website Ad
- B ○ Friend who has already participated
- C ○ Professor/CTools

Submit

[Help]

Appendix B. Instructions for Trait Questionnaire

First Screen

Welcome:

Thank you for taking the time to participate in today's experimental session. It is being conducted by the Computational and Cognitive Neuroscience Laboratory, in the Department of Psychology at the University of Michigan.

Your participation will contribute important data to our scientific studies of higher cognitive functions such as memory, decision making, and reasoning.

You are about to start the "Questionnaire" phase of the session. For the next 30-40 minutes, we would like you to fill out an important questionnaire by responding to questions that appear on your computer screen using your mouse and keyboard.

Second Screen

All of the questions on the questionnaire have been approved by the Institutional Review Board at the University of Michigan. Our methods for secure data storage have also been approved by this Board. It is critical for the scientific integrity of our experiment that you respond to the questions honestly, and accurately. You are not required to respond to any question that you are not comfortable answering. Please note, however, that your responses to the questions are confidential and anonymous. Your responses will be associated only with your anonymous Participant Code and *not* your name. Given the anonymity and confidentiality of the questionnaire, we hope that you will feel comfortable responding openly and honestly for the purposes of scientific discovery.

Appendix C. Instructions for Iowa Gambling Task

Instructions for CARD TASK

1. In front of you on the screen, there are 4 decks of cards: A, B, C, and D.
2. Once you start the game, you will select one card at a time by clicking on any deck you choose with your mouse.
3. Each time you select a card, you will win some money. Every time you win, the green bar labeled *Cash Pile* gets bigger.
4. Every so often, when you click on a card, you will win some money as usual, but then you will find out that you have lost some money as well. You will learn as you play the game how much you will win and lose. Every time you lose money, your *Cash Pile* gets smaller.
5. You are absolutely free to switch from one deck to another at any time, and as often as you wish.
6. The goal of the game is to win as much money as possible and avoid losing as much money as possible.
7. You won't know when the game will end. Simply keep on playing until the game stops.
8. We are giving you $2000 of credit, the green bar, to start the game. The red bar labeled *Borrow* is a reminder of how much money you borrowed to play the game, and how much money you have to pay back at the end of the game to figure out how much you won or lost in total. If necessary, you will be given more money. The red *Borrow* bar will show you how much you have borrowed.
9. At the end of the game, we will calculate your game result by subtracting your *Borrow* amount from your *Cash Pile* amount. The money can earn in the game is shown in the table below.

Game Result	Payment
>$4000	$50.00
$3000 to $3999	$25.00
$2000 to $2999	$12.00
$1000 to $1999	$11.00
$0 to $999	$10.00
-$1000 to -$1	$9.00
-$2000 to -$1001	$8.00
-$3000 to -$2001	$7.00
-$4000 to -$3001	$6.00
< -$4000	$5.00

10. The only hint we can give you, and the most important thing to note is this: Out of these four decks of cards, there are some that are worse than others, and to win you should try to stay away from bad decks. No matter how much you find yourself losing, you can still win the game if you avoid the worst decks.
11. Also note that the computer does not change the order of the cards once you begin the game. It does not make you lose at random, or make you lose based on the last card you chose. The amounts you can win shown in the table are all possible in this game.
12. At this time, please make sure your headphones are on. If you do not hear any sound after you click the first card, please let the experimenter know immediately.

If you have any questions, please ask the Experimenter now.

When you are ready to start the game, press the "s" key and begin choosing cards!

Appendix D. Instructions for Other Cognitive Tasks

Wisconsin Cart Sorting Task

You are about to take part in a card sorting task in which you will try to categorize cards based on a specific rule. You will see four piles of cards. You will be shown a series of cards and your goal will be to put each card into the correct pile. Press the 'c' key to continue.

Each pile has card with objects that have a different number, color, and shape. For each new card you see, your goal is to determine which pile it belongs to. You will press the 1,2 3 and 4 keys ALONG THE TOP of the keyboard to place a card into one of the four piles. Press the 'c' key to continue."

The correct pile in which to place a card depends on a sorting rule that you will have to figure out. For example, if you think the sorting rule is COLOR, then you should put blue cards in the blue pile, red cards in the red pile, etc. If you think the sorting rule is SHAPE, then you should put cards with triangles in the triangle pile, cards with circles in the circle pile, etc. If you think the sorting rule is NUMBER, then you should place cards with two objects in the pile that has two objects, cards with three objects in the pile that has three objects, etc. Press the 'c' key to continue.

As you place cards in piles, you will be shown whether your choice is CORRECT or INCORRECT. This feedback will help you figure out the correct sorting rule. Periodically, as you continue to do the task, the sorting rule will change. When it does, try to figure out the new rule as quickly as possible. You may have to change the way you place the cards in order to figure out the new rule.

If you have any questions about the task, please ask the experimenter now. Press any the 's' to start the task.

Digit Span Task

Please make sure you are wearing your headphones. You are about to take part in a memory test. You will hear and see a list of digits, one at a time. Your goal is to remember these digits in the exact order in which they were presented. After the list of digits has been presented, you will see a blank list where you can type the digits, in order, using the number keys AT THE TOP of the keyboard

If you cannot remember a digit, you can skip over it by typing the '-' key (next to the zero key). If you need to correct your response, use the 'BACKSPACE' key. When you have finished typing the list of digits, press the 'ENTER' key.

The task will begin by presenting you with three lists, each containing three digits. If you are able to correctly recall two out of the three lists, you will then move on to lists of four digits, then five digits, and so on. If you have any questions, please ask the Experimenter now. Press the 's' key to start the task.

Appendix E. Demographic and Lifestyle Questionnaire

1. **Date of birth (month/day/year, e.g. 01/12/84):**
2. **Sex:**
   A Male
   B Female
3. **Ethnicity:**
   A Hispanic/Latino
   B Not Hispanic/Latino
4. **Racial category (please indicate the one option that best describes you):**
   A American Indian/Alaska Native
   B Asian
   C Pacific Islander/Native Hawaiian
   D Black/African-American
   E White
5. **Marital status:**
   A Married
   B Divorced
   C Single
6. **Level of education (if currently in school/dropped out of school, please indicate according to current grade level/grade level reached):**
   A Never attended school
   B High School or below
   C College
   D Graduate School (Advanced Degree) or Higher
7. **Handedness:**
   A Left-handed
   B Right-handed
   C Ambidextrous
8. **Cigarette Use:**
   A I used to smoke heavily but have now quit
   B I smoke a pack or more per day
   C I smoke regularly but less than a pack per day
   D I smoke occasionally but not regularly
   E I almost never smoke cigarettes
   F I have never smoked a cigarette.
9. **Alcohol consumption:**
   A I drink very frequently (several drinks per day)
   B I drink often (a drink per day)
   C I drink somewhat frequently (a few drinks per week)
   D I drink occasionally (once every week or so)
   E I almost never drink alcohol.
   F I have never tried alcohol.
10. **Marijuana or other drug use, for recreational purposes:**
    A I used to smoke/use drugs frequently, but have now quit
    B I smoke/use drugs daily
    C I smoke/use drugs often, but not daily
    D I smoke/use drugs occasionally, but not often
    E I almost never smoke/use drugs
    F I have never tried marijuana or other drugs for recreational purposes.
11. **Gambling (e.g. sports games, horse racing, card games, casinos, etc):**
    A I used to gamble frequently but quit
    B I gamble daily
    C I gamble often, but not daily
    D I gamble occasionally, but not often
    E I almost never gamble
    F I have never gambled
12. **Amount of sleep per night, on average:**
    A 0-2 hours
    B 2-4 hours
    C 4-6 hours
    D 6-8 hours
    E More than 8 hours
13. **Math skills (please indicate how you best identify your abilities in math, on the following scale:**
    I consider myself:
    A Very Good at Math
    B Okay at Math
    C Not Very Good at Math
    D Terrible at Math

Appendix F. Internal-Reliability Consistencies

Internal-Consistency Reliability (Cronbach's Alpha)	Dissertation Study	Published Norm	Source
Compulsiveness (CI)			
Indecision & Double Checking (IDC)	0.73	0.89	(Fischer, et al., 2007)
Order & Regularity (OR)	0.61	0.88	(Fischer, et al., 2007)
Detail & Perfection (DP)	0.83	0.85	(Frost, et al., 1990)
Perfectionism (FMPS)			
Concern Over Mistakes (CM)	0.89	0.88	(Frost, et al., 1990)
Personal Standards (PS)	0.78	0.83	(Frost, et al., 1990)
Parental Expectations (PE)	0.81	0.84	(Frost, et al., 1990)
Parental Criticism (PC)	0.78	0.84	(Frost, et al., 1990)
Doubts About Actions (DA)	0.80	0.77	(Frost, et al., 1990)
Organization (ORG)	0.94	0.93	(Nenkov, Morrin, Ward, Schwartz, & Hulland, 2008)
Regret & Maximization (RMS)			
Regret (R)	0.81	.76	(Schwartz, et al., 2002)
Maximizing (M)	0.69	.70	(Frost, et al., 1990)
Indecisiveness (FIS)			
Indecisiveness (IS)	0.83	0.87	(Zuckerman, Eysenck, & Eysenck, 1978)
Sensation Seeking (SSS)			
Thrill & Adventure Seeking (TAS)	0.76	0.77	(Zuckerman, et al., 1978)
Experience Seeking (ES)	0.60	0.61	(Zuckerman, et al., 1978)
Disinhibition (DIS)	0.75	0.75	(Zuckerman, et al., 1978)
Boredom Susceptibility (BS)	0.54	0.57	(Weber, et al., 2002)
Risk Taking (DOSPERT)			
Financial (FIN)	0.76	0.88	(Weber, et al., 2002)
Health & Safety (HS)	0.67	0.88	(Weber, et al., 2002)
Recreation (REC)	0.83	0.88	(Weber, et al., 2002)
Ethical (ETH)	0.81	0.88	(Weber, et al., 2002)
Social (SOC)	0.69	0.88	(Whiteside, et al., 2005)
Impulsivity (UPPS)			
Urgency (URG)	0.88	0.89	(Whiteside, et al., 2005)
Premeditation (PRE)	0.88	0.87	(Whiteside, et al., 2005)
Perseverance (PER)	0.87	0.83	(Whiteside, et al., 2005)
Sensation Seeking (SEN)	0.89	0.85	(Stanford, et al., 2009)
Impulsivity (BIS11)			
Attention (ATT)	0.59	0.74	(Stanford, et al., 2009)
Motor (MOT)	0.72	0.59	(Stanford, et al., 2009)
Non-Planning (NPL)	0.70	0.72	(Dickman, 1990)
Impulsivity (DII)			
Functional Impulsivity (FI)	0.60	0.83	(Dickman, 1990)
Dysfunctional Impulsivity (DI)	0.65	0.86	(Carver & White, 1994)
Inhibition & Activation (BIS-BAS)			
Inhibition (BIS)	0.80	0.74	(Carver & White, 1994)
Drive (DRV)	0.71	0.76	(Carver & White, 1994)
Fun Seeking (FS)	0.69	0.66	(Carver & White, 1994)
Reward Responsivity (RR)	0.66	0.73	(Watson, et al., 1988)
General Affect (PANAS)			
Positive Affect (PA)	0.85	0.88	(Watson, et al., 1988)
Negative Affect (NA)	0.88	0.87	(Watson & Clark, 1992)
General Temperament (GTS)			
Negative Temperament (NT)	0.90	0.90 – 0.92	(Watson & Clark, 1992)
Positive Temperament (PT)	0.87	0.81 - 0.89	(Watson & Clark, 1992)
Disinhibition (DI)	0.84	0.81 – 0.86	(Srivastava, John, Gosling, & Potter, 2003)
General Traits (BFI)			
Openness (O)	0.79	0.81	(Srivastava, et al., 2003)
Conscientiousness (C)	0.83	0.82	(Srivastava, et al., 2003)
Extraversion (E)	0.90	0.88	(Srivastava, et al., 2003)
Agreeableness (A)	0.75	0.79	(Srivastava, et al., 2003)
Neuroticism (N)	0.84	0.84	(Beck, et al., 1961)
Depression (BDI)	0.78	0.86	

Appendix G. Loadings for Factor Analysis of Trait Measures

	Component			
	1	2	3	4
FMPS_CM	.747	.010	-.060	-.094
FMPS_DA	.682	.191	-.276	-.055
RMS_R	.657	.027	-.305	.109
GTS_NT	.645	.209	-.445	-.094
RMS_M	.594	.057	-.132	.249
FMPS_PS	.590	-.283	.289	.168
BISBAS_BIS	.567	-.117	-.450	-.109
FMPS_PE	.555	-.005	.117	.120
BFI_N	.554	.141	-.519	-.136
CI_OR	.517	-.127	.077	-.136
FMPS_ORG	.515	-.398	.177	-.117
FMPS_PC	.490	.126	.146	.053
UPPS_URG	.480	.472	-.056	.221
CI_DP	.390	-.336	-.081	-.171
BFI_C	-.018	-.779	.317	-.067
DII_DI	.049	.714	.105	.147
BIS11_NPL	-.206	.702	.134	.164
BIS11_ATT	.077	.675	-.285	-.031
GTS_DI	.092	.668	.116	.491
BIS11_MOT	.322	.631	.040	.351
UPPS_PER	.205	-.615	.437	.062
UPPS_PRE	.184	-.539	-.311	-.253
BFI_A	-.329	-.381	-.177	.189
BFI_E	.029	.129	.736	.056
DII_FI	-.100	.160	.723	.099
GTS_PT	.202	-.011	.678	.118
PANAS_PA	.118	-.153	.678	.036
FIS_IS	.292	.321	-.626	-.077
CI_IDC	.436	.053	-.560	-.140
BISBAS_BAS_DV	.150	.036	.531	.269
PANAS_NA	.446	.295	-.459	-.137
DOSPERT_SOC	-.121	.225	.409	.340
BFI_O	-.054	.030	.404	.103
BDI_BDI	.160	-.001	-.288	.121
UPPS_SEN	.014	.017	.261	.807
DOSPERT_REC	-.242	-.078	.114	.800
SSS_TAS	.055	-.122	.113	.744
BISBAS_BAS_FS	.063	.327	.211	.681
DOSPERT_HS	-.189	.326	.040	.577
DOSPERT_ETH	-.034	.408	.051	.561
SSS_DIS	.239	.374	.075	.489
DOSPERT_FIN	.098	.171	-.037	.463
SSS_ES	-.312	.209	.174	.412
BISBAS_BAS_RR	.350	-.072	.002	.357
SSS_BS	.064	.326	.231	.331

REFERENCES

Anderson, J. R. (1991). The Adaptive Nature of Human Categorization. *Psychological review, 98*(3), 409-429.

Appollonio, I. M., Russo, A., Isella, V., Forapani, E., Villa, M. L., Piolti, R., et al. (2002, Oct 23-25). *Cognitive estimation: comparison of two tests in nondemented parkinsonian patients.* Paper presented at the 29th National Congress on Parkinson Parkinsonism Dementia, Lecce, Italy.

Ashby, E. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149-178.

Avogadri, R., & Valentini, G. (2007, Jul 07-10). *Fuzzy ensemble clustering for DNA Microarray data analysis.* Paper presented at the 7th International Workshop on Fuzzy Logic and Applications, Camogli, Italy.

Barratt, E. S. (1985). Impulsiveness subtraits: Arousal and information processing. In J. T. Spence & C. E. Izard (Eds.), *Motivation, Emotion and Personality* (pp. 137-146). North Holland: Elsevier Science.

Bechara, A. (2004). The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. *Brain and cognition, 55*(1), 30-40.

Bechara, A. (2007). *Iowa Gambling Task Professional Manual*. Lutz: PAR, Inc.

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior, 52*(2), 336-372.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences dollowing damage to human prefrontal cortex. *Cognition, 50*(1-3), 7-15.

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex, 10*(3), 295-307.

Bechara, A., Damasio, H., & Damasio, A. R. (2003). Role of the amygdala in decision-making. *Annals of the New York Academy of Sciences, 985*, 356-369.

Bechara, A., Damasio, H., Damasio, A. R., & Lee, G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of Neuroscience, 19*(13), 5473-5481.

Bechara, A., Damasio, H., Tranel, D., & Anderson, S. W. (1998). Dissociation of working memory from decision making within the human prefrontal cortex. *Journal of Neuroscience, 18*(1), 428-437.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*(5304), 1293-1295.

Bechara, A., Dolan, S., Denburg, N., Hindes, A., Anderson, S. W., & Nathan, P. E. (2001). Decision-making deficits, linked to a dysfunctional ventromedial prefrontal cortex, revealed in alcohol and stimulant abusers. *Neuropsychologia, 39*(4), 376-389.

Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain, 123*, 2189-2202.

Bechara, A., Tranel, D., Damasio, H., & Damasio, A. R. (1996). Failure to respond autonomically to anticipated future outcomes following damage to prefrontal cortex. *Cerebral Cortex, 6*(2), 215-225.

Beck, A. T., Erbaugh, J., Ward, C. H., Mock, J., & Mendelsohn, M. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*(6), 53-63.

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories-IA and -II in psychiatric outpatients. *Journal of Personality Assessment, 67*(3), 588-597.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation.

Blais, A. R., & Weber, E. U. (2006). A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. *Judgment and Decision Making Journal, 1*(1), 33-47.

Bolla, K. I., Eldreth, D. A., Matochik, J. A., & Cadet, J. L. (2004). Sex-related differences in a gambling task and its neurological correlates. *Cerebral Cortex, 14*(11), 1226-1232.

Bolla, K. I., Eldreth, D. A., Matochik, J. A., & Cadet, J. L. (2005). Neural substrates of faulty decision-making in abstinent marijuana users. *Neuroimage, 26*(2), 480-492.

Brand, M., & Altstotter-Gleich, C. (2008). Personality and decision-making in laboratory gambling tasks - Evidence for a relationship between deciding advantageously under risk conditions and perfectionism. *Personality and Individual Differences, 45*(3), 226-231.

Brand, M., Recknor, E. C., Grabenhorst, F., & Bechara, A. (2007). Decisions under ambiguity and decisions under risk: Correlations with executive functions and comparisons of two different gambling tasks with implicit and explicit rules. *Journal of Clinical and Experimental Neuropsychology, 29*(1), 86-99.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York,: Wiley.

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the bechara gambling task. *Psychological Assessment, 14*(3), 253-262.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3*(1), 1-27.

Campbell, M. C., Stout, J. C., & Finn, P. R. (2004). Reduced autonomic responsiveness to gambling task losses in Huntington's disease. *Journal of International Neuropsychological Society, 10*(2), 239-245.

Carver, C. S., & White, T. L. (1994). Behavioral-inhibition,behavioral-activation, and affective responses to impending reward and punishment - The BIS BAS scales. *Journal of Personality and Social Psychology, 67*(2), 319-333.

Colombetti, G. (2008). The somatic marker hypotheses, and what the Iowa Gambling Task does and does not show. *British Journal for the Philosophy of Science, 59*(1), 51-71.

Corrado, G. S., Sugrue, L. P., Seung, H. S., & Newsome, W. T. (2005). Linear-Nonlinear-Poisson models of primate choice dynamic. *Journal of the experimental analysis of behavior, 84*(3), 581-617.

Crone, E. A., Bunge, S. A., Latenstein, H., & van der Molen, M. W. (2005). Characterization of children's decision making: Sensitivity to punishment frequency, not task complexity. *Child Neuropsychology, 11*(3), 245-263.

Crone, E. A., & van der Molen, M. W. (2004). Developmental changes in real life decision making: Performance on a gambling task previously shown to depend on the ventromedial prefrontal cortex. *Developmental Neuropsychology, 25*(3), 251-279.

Crone, E. A., & van der Molen, M. W. (2007). Development of decision making in school-aged children and adolescents: Evidence from heart rate and skin conductance analysis. *Child Development, 78*(4), 1288-1301.

Crone, E. A., Vendel, I., & van der Molen, M. W. (2003). Decision-making in disinhibited adolescents and adults: insensitivity to future consequences or driven by immediate reward? *Personality and Individual Differences, 35*(7), 1625-1641.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: G.P. Putnam.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 351*(1346), 1413-1420.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 1*(2), 224-227.

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology, 16*(2), 199-204.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441*(7095), 876-879.

Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: Massachusetts Institute of Technology Press.

Della Sala, S., MacPherson, S. E., Phillips, L. H., Sacco, L., & Spinnler, H. (2001, Mar 09). *How many camels are there in Italy? Cognitive estimates standardised on the Italian population.* Paper presented at the Meeting of the Italian Neurology Society, Rimini, Italy.

Denburg, N. L., Weller, J. A., Yamada, T. H., Shivapour, D. M., Kaup, A. R., LaLoggia, A., et al. (2009). Poor decision making among older adults is related to elevated levels of neuroticism. *Annals of Behavioral Medicine, 37*(2), 164-172.

Desmeules, R., Bechara, A., & Dube, L. (2008). Subjective valuation and asymmetrical motivational systems: implications of scope insensitivity for decision making. *Journal of Behavioral Decision Making, 21*(2), 211-224.

Dickman, S. J. (1990). Functional and dysfunctional impulsivity - Personality and cognitive correlates. *Journal of Personality and Social Psychology, 58*(1), 95-102.

Duarte, F. J., Fred, A. L. N., Lourenco, A., & Rodrigues, M. F. (2005). Weighting cluster ensembles in evidence accumulation clustering. *2005 Portuguese Conference on Artificial Intelligence, Proceedings*, 159-167.

Dubes, R., & Jain, A. K. (1976). Clustering techniques: The user's dilemma. *Pattern Recognition, 8*(4), 247-260.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). New York: Wiley.

Dudoit, S., & Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology, 3*(7), 1-21.

Dudoit, S., & Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics, 19*(9), 1090-1099.

Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association, 97*(457), 77-87.

Dunn, B. D., Dalgleish, T., & Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience and Biobehavioral Reviews, 30*(2), 239-271.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Cybernetics and Systems: An International Journal, 4*(1), 95 - 104.

Ellsberg, D. (1961). Risk, ambiguity and the savage axioms. *Quarterly Journal of Economics, 75*(4), 643-669.

Fellows, L. K. (2007). The role of orbitofrontal cortex in decision making - A component process account *Linking Affect to Action: Critical Contributions of the Orbitofrontal Cortex* (Vol. 1121, pp. 421-430). Oxford: Blackwell Publishing.

Fellows, L. K., & Farah, M. J. (2003). Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain, 126*, 1830-1837.

Fellows, L. K., & Farah, M. J. (2005). Different underlying impairments in decision-making following ventromedial and dorsolateral frontal lobe damage in humans. *Cerebral Cortex, 15*(1), 58-63.

Fischer, J., Corcoran, K., & Corcoran, K. (2007). *Measures for Clinical Practice and Research: A Sourcebook* (4th ed.). Oxford: Oxford University Press.

Franken, I. H. A., & Muris, P. (2005). Individual differences in decision-making. *Personality and Individual Differences, 39*(5), 991-998.

Franken, I. H. A., van Strien, J. W., Nijs, I., & Muris, P. (2008). Impulsivity is associated with behavioral decision-making deficits. *Psychiatry Research, 158*(2), 155-163.

Fred, A. L. N., & Jain, A. K. (2003, Jun 18-20). *Robust data clustering.* Paper presented at the Conference on Computer Vision and Pattern Recognition, Madison, Wi.

Fred, A. L. N., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(6), 835-850.

Fred, A. L. N., & Jain, A. K. (2006). Learning pairwise similarity for data clustering. *18th International Conference on Pattern Recognition, Vol 1, Proceedings*, 925-928.

Frost, R. O., Marten, P., Lahart, C., & Rosenblate, R. (1990). The dimensions of perfectionism *Cognitive Therapy and Research, 14*(5), 449-468.

Frost, R. O., & Shows, D. L. (1993). The nature and measure of compulsive indecisiveness. *Behaviour Research and Therapy, 31*(7), 683-692.

Fukui, H., Murai, T., Fukuyama, H., Hayashi, T., & Hanakawa, T. (2005). Functional activity related to risk anticipation during performance of the Iowa gambling task. *Neuroimage, 24*(1), 253-259.

Garon, N., & Moore, C. (2004). Complex decision-making in early childhood. *Brain and Cognition, 55*(1), 158-170.

Garon, N., & Moore, C. (2007). Awareness and symbol use improves future-oriented decision making in preschoolers. *Developmental Neuropsychology, 31*(1), 39-59.

Glimcher, P. W. (2008). *Neuroeconomics: Decision Making and the Brain*. London: Academic.

Goudriaan, A. E., Grekin, E. R., & Sher, K. J. (2007). Decision making and binge drinking: A longitudinal study. *Alcoholism-Clinical and Experimental Research, 31*(6), 928-938.

Grant, D. A., & Berg, E. A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology, 38*(4), 404-411.

Greve, K. W., Stickle, T. R., Love, J. A., Bianchini, K. J., & Stanford, M. S. (2003, Feb 05-08). *Latent structure of the Wisconsin Card Sorting Test: a confirmatory factor analytic study.* Paper presented at the 31st Annual Meeting of the International Neuropsychological Society, Honolulu, HI.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Clustering validity checking methods: Part II. *Sigmod Record, 31*(3), 19-27.

Halkidi, M., & Vazirgiannis, M. (2001). *Clustering validity assessment: Finding the optimal partitioning of a data set.* Paper presented at the IEEE International Conference on Data Mining, San Jose, CA.

Harmsen, H., Bischof, G., Brooks, A., Hohagen, F., & Rumpf, H. J. (2006). The relationship between impaired decision-making, sensation seeking and readiness to change in cigarette smokers. *Addictive Behaviors, 31*(4), 581-592.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics, 28*(1), 100-108.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction*. New York: Springer.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior, 4*, 267-272.

Hinson, J. M., Jameson, T. L., & Whitney, P. (2002). Somatic markers, working memory, and decision making. *Cognitive, Affective, & Behavioral Neuroscience, 2*, 341-353.

Hinson, J. M., Whitney, P., Holben, H., & Wirick, A. K. (2006). Affective biasing of choices in gambling task decision making. *Cognitive Affective & Behavioral Neuroscience, 6*(3), 190-200.

Hooper, C. J., Luciana, M., Conklin, H. M., & Yarger, R. S. (2004). Adolescents' performance on the Iowa gambling task: Implications for the development of decision making and ventromedial prefrontal cortex. *Developmental Psychology, 40*(6), 1148-1158.

Hooper, C. J., Luciana, M., Wahlstrom, D., Conklin, H. M., & Yarger, R. S. (2008). Personality correlates of Iowa Gambling Task performance in healthy adolescents. *Personality and Individual Differences, 44*(3), 598-609.

Huizenga, H. M., Crone, E. A., & Jansen, B. J. (2007). Decision-making in healthy children, adolescents and adults explained by the use of increasingly complex proportional reasoning rules. *Developmental Science, 10*(6), 814-825.

Jameson, T. L., Hinson, J. M., & Whitney, P. (2004). Components of working memory and somatic markers in decision making. *Psychonomic Bulletin & Review, 11*(3), 515-520.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The Big Five Inventory-Versions 4a and 54. Berkely, CA: Institute of Personality and Social Research, University of California.

Kagan, D. M., & Squires, R. L. (1985). Measuring nonpathological compulsiveness. *Psychological Reports, 57*(2), 559-563.

Kahneman, D., & Tversky, A. (1979). Prospect theory - analysis of decision under risk. *Econometrica, 47*(2), 263-291.

Kahneman, D., & Tversky, A. (2000). *Choices, Values, and Frames*. Cambridge, UK: Cambridge University Press.

Kakade, S., & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks, 15*(4-6), 549-559.

Kalidindi, K., & Bowman, H. (2007). Using epsilon-greedy reinforcement learning methods to further understand ventromedial prefrontal patients' deficits on the Iowa Gambling Task. *Neural Networks, 20*(6), 676-689.

Kerr, A., & Zelazo, P. D. (2004). Development of "hot" executive function: The children's gambling task. *Brain and Cognition, 55*(1), 148-157.

Krzanowski, W. J., & Lai, Y. T. (1988). A Criterion for Determining the Number of Groups in a Data Set Using Sum-of-Squares Clustering. *Biometrics, 44*(1), 23-34.

Lakey, C. E., Rose, P., Campbell, W. K., & Goodie, A. S. (2008). Probing the link between narcissism and gambling: The mediating role of judgment and decision-making biases. *Journal of Behavioral Decision Making, 21*(2), 113-137.

Lamberts, K., & Shanks, D. R. (1997). *Knowledge, concepts and categories*. Hove, East Sussex, UK: Psychology Press.

Lane, S. D., Yechiam, E., & Busemeyer, J. R. (2006). Application of a computational decision model to examine acute drug effects on human risk taking. *Experimental and Clinical Psychopharmacology, 14*(2), 254-264.

Lange, T., Roth, V., Braun, M. L., & Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural Computation, 16*(6), 1299-1323.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior, 84*(3), 555-579.

Law, M. H. C., Topchy, A. P., & Jain, A. K. (2004, Jun 27-Jul 02). *Multiobjective data clustering*. Paper presented at the Conference on Computer Vision and Pattern Recognition, Washington, DC.

Lawrence, N. S., Jollant, F., O'Daly, O., Zelaya, F., & Phillips, M. L. (2009). Distinct roles of prefrontal cortical subregions in the Iowa Gambling Task. *Cerebral Cortex, 19*(5), 1134-1143.

Lilliefor, H. W. (1967). On Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62*(318), 399-&.

Lourenco, A., & Fred, A. (2005). Ensemble methods in the clustering of string patterns. *WACV 2005: Seventh IEEE Workshop on Applications of Computer Vision, Proceedings*, 143-148.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review, 66*(2), 81-95.

Luce, R. D. (1977). Choice axiom after 20 years. *Journal of Mathematical Psychology, 15*(3), 215-233.

Luo, F., & Liu, J. (2007, Sep 21-23). *Clustering analysis of microarray gene expression data with new clustering ensemble method.* Paper presented at the 2nd International Symposium on Intelligence Computation and Application (ISICA 2007), Wuhan, Peoples Republic of China.

Maia, T. V., & McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: What participants really know in the Iowa gambling task. *Proceedings of the National Academy of Sciences of the United States of America, 101*(45), 16075-16080.

Masulli, F., & Rovetta, S. (2003, Jul 20-24). *An ensemble approach to variable selection for classification of DNA microarray data.* Paper presented at the International Joint Conference on Neural Networks, Portland, OR.

Medin, D. L., & Smith, E. E. (1984). Concepts and Concept-Formation. *Annual Review of Psychology, 35*, 113-138.

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data. *Psychometrika, 50*(2), 159-179.

Mimura, M., Oeda, R., & Kawamura, M. (2006). Impaired decision-making in Parkinson's disease. *Parkinsonism and Related Disorders, 12*(3), 169-175.

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature, 377*(6551), 725-728.

Mueller, S. T. (2008). PEBL: The psychology experiment building language (Version 0.09).

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review, 92*(3), 289-316.

Nenkov, G. Y., Morrin, M., Ward, A., Schwartz, B., & Hulland, J. (2008). A short form of the Maximization Scale: Factor structure, reliability and validity studies. *Judgment and Decision Making Journal, 3*(5), 371-U371.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001, Dec 03-08). *On spectral clustering: Analysis and an algorithm.* Paper presented at the 15th Annual Conference on Neural Information Processing Systems (NIPS), Vancouver, Canada.

Niv, Y., Daw, N. D., & Dayan, P. (2006). Choice values. *Nature Neuroscience, 9*(8), 987-988.

O'Carroll, R. E., & Papps, B. P. (2003). Decision making in humans: the effect of manipulating the central noradrenergic system. *Journal of Neurology Neurosurgery and Psychiatry, 74*(3), 376-378.

O'Doherty, J., Rolls, E. T., Francis, S., Bowtell, R., & McGlone, F. (2001). Representation of pleasant and aversive taste in the human brain. *Journal of Neurophysiology, 85*(3), 1315-1321.

O'Doherty, J. P., Critchley, H., Deichmann, R., & Dolan, R. J. (2003). Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. *Journal of Neuroscience, 23*(21), 7931-7939.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron, 38*(2), 329-337.

O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304*(5669), 452-454.

Osherson, D. N., Wilkie, O., Smith, E. E., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*(2), 185-200.

Overman, W. H. (2004). Sex differences in early childhood, adolescence, and adulthood on cognitive tasks that rely on orbital prefrontal cortex. *Brain and Cognition, 55*(1), 134-147.

Oya, H., Adolphs, R., Kawasaki, H., Bechara, A., Damasio, A., & Howard, M. A. (2005). Electrophysiological correlates of reward prediction error recorded in the human prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 102*(23), 8351-8356.

Pagonabarraga, J., Garcia-Sanchez, C., Llebaria, G., Pascual-Sedano, B., Gironell, A., & Kulisevsky, J. (2007). Controlled study of decision-making and cognitive impairment in Parkinson's disease. *Movement Disorders, 22*(10), 1430-1435.

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt Impulsiveness Scale. *Journal of Clinical Psychology, 51*(6), 768-774.

Preston, S. D., Buchanan, T. W., Stansfield, R. B., & Bechara, A. (2007). Effects of anticipatory stress on decision making in a gambling task. *Behavioral Neuroscience, 121*(2), 257-263.

Qiu, P., Wang, Z. J., & Liu, K. J. R. (2005). Ensemble dependence model for classification and prediction of cancer and normal gene expression data. *Bioinformatics, 21*(14), 3114-3121.

Reavis, R., & Overman, W. H. (2001). Adult sex differences on a decision-making task previously shown to depend on the orbital prefrontal cortex. *Behavioral Neuroscience, 115*(1), 196-206.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*. New York: J. Wiley.

Rolls, E. T. (1996). The orbitofrontal cortex. *Philosophical transactions of the Royal Society of London. Series B: Biological sciences, 351*(1346), 1433-1443; discussion 1443-1434.

Rolls, E. T. (2004). The functions of the orbitofrontal cortex. *Brain and Cognition, 55*(1), 11-29.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80*(1), 1-27.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology, 57*, 87-115.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593-1599.

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex, 10*(3), 272-283.

Schutter, D., de Haan, E. H. F., & van Honk, J. (2004). Anterior asymmetrical alpha activity predicts Iowa gambling performance: distinctly but reversed. *Neuropsychologia, 42*(7), 939-943.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology, 83*(5), 1178-1197.

Shallice, T., & Evans, M. E. (1978). Involvement of the frontal lobes in cognitive estimation. *Cortex, 14*(2), 294-303.

Shizgal, P. (1997). Neural basis of utility estimation. *Current Opinion in Neurobiology, 7*(2), 198-208.

Shurman, B., Horan, W. P., & Ntuechterlein, K. H. (2005). Schizophrenia patients demonstrate a distinctive pattern of decision-making impairment on the Iowa Gambling Task. *Schizophrenia Research, 72*(2-3), 215-224.

Singh, S., Lewis, R. L., & Barto, A. G. (2009). *Where do rewards come from?* Paper presented at the Proceedings of the 31th Annual Conference of the Cognitive Science Society, Amsterdam, Netherlands.

Smyth, C., & Coomans, D. (2007, Apr 01-05). *Clustering microarrays with predictive weighted ensembles.* Paper presented at the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, Honolulu, HI.

Spreen, O., & Strauss, E. (1998a). *A compendium of neuropsychological tests : administration, norms, and commentary* (2nd ed.). New York: Oxford University Press.

Spreen, O., & Strauss, E. (1998b). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary* (2nd ed.). New York: Oxford University Press.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology, 84*(5), 1041-1053.

Stanford, M. S., Mathias, C. W., Dougherty, D. M., Lake, S. L., Anderson, N. E., & Patton, J. H. (2009). Fifty years of the Barratt Impulsiveness Scale: An update and review. *Personality and Individual Differences, 47*(5), 385-395.

Stout, J. C., Busemeyer, J., Bechara, A., & Lin, A. (2002). Cognitive modeling of decision making in a simulated gambling task in frontal or somatosensory cortex damage. *Journal of Cognitive Neuroscience*, 75-75.

Stout, J. C., Busemeyer, J. R., Lin, A. L., Grant, S. J., & Bonson, K. R. (2004). Cognitive modeling analysis of decision-making processes in cocaine abusers. *Psychonomic Bulletin & Review, 11*(4), 742-747.

Stout, J. C., Rodawalt, W. C., & Siemers, E. R. (2001). Risky decision making in Huntington's disease. *Journal of the International Neuropsychological Society, 7*(1), 92-101.

Suhr, J. A., & Tsanadis, J. (2006). Affect and personality correlates of the Iowa Gambling Task. *Personality and Individual Differences, 43*(1), 27-36.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Sweitzer, M. M., Allen, P. A., & Kaut, K. P. (2008). Relation of individual differences in impulsivity to nonclinical emotional decision making. *Journal of the International Neuropsychological Society, 14*(5), 878-882.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics*. Boston: Allyn and Bacon.

Thut, G., Schultz, W., Roelcke, U., Nienhusmeier, M., Missimer, J., Maguire, R. P., et al. (1997). Activation of the human brain by monetary reward. *Neuroreport, 8*(5), 1225-1228.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 63*, 411-423.

Tomb, I., Hauser, M., Deldin, P., & Caramazza, A. (2002). Do somatic markers mediate decisions on the gambling task? *Nature Neuroscience, 5*(11), 1103-1104.

Topchy, A. P., Law, M. H. C., Jain, A. K., & Fred, A. L. (2004, Nov 01-04). *Analysis of consensus partition in cluster ensemble.* Paper presented at the 4th IEEE International Conference on Data Mining, Brighton, England.

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature, 398*(6729), 704-708.

Turnbull, O. H., Berry, H., & Bowman, C. H. (2003). Direct versus indirect emotional consequences on the Iowa Gambling Task. *Brain and Cognition, 53*(2), 389-392.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453-458.

van den Bos, R., Houx, B. B., & Spruijt, B. A. (2006). The effect of reward magnitude differences on choosing disadvantageous decks in the Iowa Gambling Task. *Biological Psychology, 71*(2), 155-161.

van Honk, J., Hermans, E. J., Putman, P., Montague, B., & Schutter, D. (2002). Defective somatic markers in sub-clinical psychopathy. *Neuroreport, 13*(8), 1025-1027.

van Honk, J., Schutter, D., Hermans, E. J., & Putman, P. (2003). Low cortisol levels and the balance between punishment sensitivity and reward dependency. *Neuroreport, 14*(15), 1993-1996.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236-&.

Watson, D., & Clark, L. A. (1992). On traits and temperament - General and specific factors of emotional experience and their relation to the 5-Factor Model. *Journal of Personality, 60*(2), 441-476.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect - the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

Weber, E. U., Blais, A. R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making, 15*(4), 263-290.

Whiteside, S. P., & Lynam, D. R. (2001). The Five Factor Model and impulsivity: using a structural model of personality to understand impulsivity. *Personality and Individual Differences, 30*(4), 669-689.

Whiteside, S. P., Lynam, D. R., Miller, J. D., & Reynolds, S. K. (2005). Validation of the UPPS impulsive behaviour scale: a four-factor model of impulsivity. *European Journal of Personality, 19*(7), 559-574.

Xiaohua, H., & Illhoi, Y. (2004). *Cluster ensemble and its applications in gene expression analysis*. Paper presented at the Proceedings of the second conference on Asia-Pacific bioinformatics - Volume 29.

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review, 12*(3), 387-402.

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science, 16*(12), 973-978.

Yu, Z. W., Wong, H. S., & Wang, H. Q. (2007). Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics, 23*(21), 2888-2896.

Zermatten, A., Van der Linden, M., d'Acremont, M., Jermann, F., & Bechara, A. (2005). Impulsivity and decision making. *Journal of Nervous and Mental Disease, 193*(10), 647-650.

Zhao, H., Liang, J. M., & Hu, H. H. (2006). *Clustering validity based on the improved Hubert Gamma statistic and the separation of clusters.* Paper presented at the 1st International Conference on Innovative Computing, Information and Control (ICICIC 2006), Beijing, Peoples Republic of China.

Zuckerman, M. (1964). Development of a sensation-seeking scale. *Journal of Consulting Psychology, 28*(6), 477-482.

Zuckerman, M. (1971). Dimensions of sensation seeking. *Journal of Consulting and Clinical Psychology, 36*(1), 45-52.

Zuckerman, M., Eysenck, S., & Eysenck, H. J. (1978). Sensation seeking in England and America: cross-cultural, age, and sex comparisons. *Journal of Consulting and Clinical Psychology, 46*(1), 139-149.

Zuckerman, M., & Link, K. (1968). Construct validity for the sensation-seeking scale. *Journal of Consulting and Clinical Psycholology, 32*(4), 420-426.